Harvard Data Science Review

Tackling COVID-19 through Responsible AI Innovation: Five Steps in the Right Direction

David Leslie

Published on: May 27, 2020

Updated on: Jun 05, 2020

ABSTRACT

Innovations in data science and artificial intelligence/machine learning (AI/ML) have a central role to play in supporting global efforts to combat COVID-19. The versatility of AI/ML technologies enables scientists and technologists to address an impressively broad range of biomedical, epidemiological, and socioeconomic challenges. This wide-reaching scientific capacity, however, also raises a diverse array of ethical challenges. The need for researchers to act quickly and globally in tackling SARS-CoV-2 demands unprecedented practices of open research and responsible data sharing at a time when innovation ecosystems are hobbled by proprietary protectionism, inequality, and a lack of public trust. Moreover, societally impactful interventions like digital contact tracing are raising fears of 'surveillance creep' and are challenging widely held commitments to privacy, autonomy, and civil liberties. Prepandemic concerns that data-driven innovations may function to reinforce entrenched dynamics of societal inequity have likewise intensified given the disparate impact of the virus on vulnerable social groups and the life-and-death consequences of biased and discriminatory public health outcomes. To address these concerns, I offer five steps that need to be taken to encourage responsible research and innovation. These provide a practice-based path to responsible AI/ML design and discovery centered on open, accountable, equitable, and democratically governed processes and products. When taken from the start, these steps will not only enhance the capacity of innovators to tackle COVID-19 responsibly, they will, more broadly, help to better equip the data science and AI/ML community to cope with future pandemics and to support a more humane, rational, and just society.

Keywords: COVID-19, AI ethics, responsible research and innovation, open science, digital contact tracing, public trust[1]

1. Introduction

In June 1955, the great Hungarian mathematician and polymath John Von Neumann published a popular essay entitled "Can we survive technology?" (Von Neumann, 1955). Von Neumann, then stricken with terminal cancer, wrote about what he called "the maturing crisis of technology," a situation in which the global effects of accelerating technological advancement were outpacing the development of ethical and political self-understandings that were capable of responsibly managing such an explosion of innovation. This crisis, he feared, was creating unprecedented dangers of species-level self-destruction ranging from geoengineering and unbridled automation to nuclear holocaust. At the same, he puzzled that "technological power, technological efficiency as such" was "an ambivalent achievement." That is, the very forces of ingenuity that were creating the dangers of anthropogenic self-annihilation contained within themselves the potential to benefit humanity. They possessed, *in potentia*, a kind of countervailing redeeming power.

As society now grapples with a different kind of crisis than the one Von Neumann had in mind, his reflections are no less relevant for thinking through data-driven technology's direction of travel in confronting the challenges presently faced. The maturing crisis of technology to which he referred applies especially to the field of artificial intelligence (AI) and machine learning (ML).[2] In less than a generation, exponential leaps in information-processing power have coalesced with the omnipresent data extraction capabilities of an ever more dynamic, integrated, and connected digital world to provide a fecund spawning ground for the explosion of AI/ML technologies. And, as these innovations have advanced apace—as the scope of their impacts has come to stretch from the most intimate depths of self-development to the fate of the biosphere itself—we need ever more to reflect soberly on Von Neumann's worry: Have we developed the novel ethical and political self-understandings, values, practices, and forms of life necessary to responsibly steer and constrain the rapid proliferation of AI/ML technologies?

By all accounts, society has, in fact, struggled to keep up. The all-too-common 'break first, think later' attitude of many of those at the wheel of commercial AI/ML innovation has been a recipe for financial success simultaneously as it has been a fast track in the race to the ethical bottom. Prominent examples of algorithmic bias and discrimination, of proprietary black boxes and organizational opacity, and of macroscale behavioral tracking, curating, and nudging have led to social consternation and distrust. More troubling still, AI/ML-enabled capabilities for hyperpersonalized targeting, anticipatory calculation, and algorithmic administration at scale have manifested in intrusive hazard-preemption regimes (O'Grady, 2015) ranging from data-driven border control (Amoore, 2009; Amoore & Raley, 2017) to predictive policing and commercial surveillance. They have also enabled emergent digital autocracies to engage in population-level individual monitoring and mass disciplinary control. Most consequentially, though, global prospects for a divisive geopolitical sprint to technological ascendency in AI/ML are now opening up new possibilities for destructive struggles of an unprecedented scale. In virtue of the accelerating pace of digital innovation propelled by the hasty pursuit of competitive advantage, such conflicts may soon pose very real dangers to the future of life itself—dangers extending from calamitous cyberattacks on infrastructural vulnerabilities, algorithmically streamlined biological warfare, and human enhancement 'arms races' to smart nuclear terrorism and the potentially genocidal proliferation of lethal autonomous weapons.

All this would seem to leave us, then, at the perilous crossroads of two crises—one rooted in the destructive potentials of our extant technological practices and another demanding that those same practices be marshalled as a saving power to combat the destruction inflicted by an inhuman biological agent. Faced with the current public health crisis, data scientists and AI/ML innovators may be inclined to ask: Are we ready for this? Can we find a responsible path to wielding our technological efficacy ethically and safely? In what follows, I claim that this crossroads need not induce paralysis as to which way we should go, and, in fact, presents us with clear signage for finding the right way forward. When pressed into the service of the public good, biomedical AI/ML applications have already made noteworthy progress in assisting doctors and researchers in the areas of diagnostics, prognostics, genomics, drug discovery, epidemiology, and mobile health monitoring.[3] And, though all of these areas of advancement hold the great promise of helping health care professionals to combat COVID-19, they also come with substantial ethical hazards. What we need now are actionable means to navigate these.

Here, I argue that, in our current dilemma, the data science and AI community, writ large, ought to draw upon the hard-gained critical leverage and normative resources provided by applied ethics, responsible research and innovation (RRI), science and technology studies (STS), and AI/ML ethics to close the gap between fleetfooted innovation and slow-moving ethical and social values. In the first two sections I will motivate this deliberate turn to responsible AI innovation. Starting with the unparalleled challenges presented to data scientists and AI/ML researchers by the pandemic, I will explore how some of the more intractable ethical issues already faced by AI/ML innovation are raising their heads in the present circumstance of a global public health crisis that is placing researchers under unprecedented pressures to rapidly respond. I will then lay out some of the ethical pitfalls and societal challenges faced by the potential introduction of digital contact tracing and health monitoring technologies into a networked, big data society that is in peril of rocking back and forth between the Scylla of mass digital surveillance and the Charybdis of an ethically chilling but privacy-securing automation-all-the-way-down. Finally, I will move on to offering five steps toward responsible AI/ML research and innovation that need to be taken to address these concerns. Drawing upon current thinking in applied AI/ML ethics, social scientific approaches to data-driven technologies, and RRI, these steps suggest a means of attaining and assessing open, accountable, equitable, and democratically governed AI/ML processes and products.

2. Combating COVID-19 on the Second Front: New Challenges, Old Problems

The tasks that face us as a society, at present, are posing extraordinary ethical challenges of a kind that many of us have never before experienced. On the frontlines of the pandemic, our health care professionals are confronted with a merciless convergence of limited resources and surging illness. They are having to make difficult life-and-death choices about who receives critically needed care and how. As they do battle in the global struggle against the pandemic, these heroes, and all the essential workers who keep the ship of society afloat in times of crisis, face grueling and unprecedented demands for self-sacrifice, moral fortitude, and resilience. These trials of conscience and character are testing the depths and shallows of us all and transforming our lives forever.

But the global struggle against COVID-19 is also being fought on a second crucial front with its own set of broad-reaching ethical challenges. While so many of our doctors, nurses, and key workers combat the virus on the frontlines, researchers and technologists too must tirelessly labor on the frontiers of biomedical, epidemiological, and societal innovation, so that their scientific discoveries can be employed to manage the spread of the virus and mitigate its effects.

In the data science and AI community, such second-front efforts are already well under way. Machine learning and data-driven technologies are already augmenting human capacities to better tackle the challenges of the pandemic (Bullock et al., 2020). These AI/ML-assisted interventions range from AI-supported radiological diagnostics (Ai et al., 2020; Gozes et al., 2020; Shan et al., 2020; Wang, Kang, et al., 2020), prognostics based on clinical data (Pourhomayoun & Shakibi, 2020; Yan et al., 2020; Qi et al., 2020), pharmaceutical discovery and repurposing (Beck et al., 2020; Hu et al., 2020), and test-kit development (Metsky et al., 2020) to methods of protein and RNA profiling that are shedding light on virus function and disease progression (Jumper et al., 2020; Senior et al., 2020; Zhang et al., 2020). Likewise, in the area of research support, vast troves of existing biomedical literature are being mined by AI/ML technologies to identify clinically established drugs and treatment methods that may be of use in fighting the SARS-CoV-2 infection (Ge et al., 2020). Taken cumulatively, such interventions are helping to greatly enhance the quality and speed of the human response to the outbreak.

In the areas of epidemiological modeling and social-demographic analysis too, the high-dimensional processing capacity of AI/ML applications are helping scientists to generate more effective real-time forecasts of the spread of infection and of the locations of potential outbreaks (Al-qaness et al., 2020; Hu et al., 2020). Signally, at the very outset of the pandemic an AI/ML system from the Canadian health monitoring platform BlueDot warned of the outbreak nearly two weeks before the World Health Organization (WHO) made its own announcement (Marks, 2020; Niiler, 2020). AI/ML-supported population-level insight is also being used to combat the dissemination of misinformation about the pandemic (the spread of the so-called infodemic) (Boberg et al., 2020; Chen et al., 2020; Cinelli et al., 2020; Mejova & Kalimeri, 2020). Better knowledge about the reach and sources of the propagation of misinformation will help to produce more effectual policy interventions and to promote more critical information consumption at scale.

From a wide-angled view, this expansive spectrum of AI-supported interventions demonstrates an unprecedented opportunity for the data science and AI community to press its energies and talents into the service of advancing the public good. And yet, the novel practical and sociotechnical challenges posed by the current coronavirus pandemic suggest that members of this broad church must proceed with a heightened sobriety and vigilance. The prevalent urgency for answers and the pains and pressures of producing highly impactful research in the context of a global health crisis are only magnifying many of the existing ethical concerns raised by the use of AI/ML in medicine, epidemiology, and public health, even in normal times.

This has already been well-illustrated in a timely review of 31 prediction models from 27 early studies of COVID-19 by Wynants et al. (2020). In their critical appraisal, the authors find these models to be "at high risk of [statistical] bias, mostly because of non-representative selection of control patients, exclusion of patients who had not experienced the event of interest by the end of the study, and high risk of model overfitting" (p. 1). This risk of bias is attributed to "poor reporting and poor methodological conduct for participation selection, predictor description, and statistical methods used" (p. 7). The review also highlights the fact that the 12 diagnostic imaging studies of CT scans at hand lacked clear information on how the data was preprocessed and presented highly complex algorithms that transformed imaging data into predictors in opaque and unintelligible ways. Though the authors acknowledge that these studies, as a whole, were "done under severe time constraints caused by urgency," they also caution that, in a highly distressed clinical environment, practitioners might be encouraged to "implement prediction models without sufficient documentation and validation," leading to potentially harmful outcomes (pp. 8–9). Each of the issues raised by Wynants et al. is worthy of some unpacking.

2.1. Pitfalls of COVID-19-Related Research I: Algorthmic Bias and Discrimination

First, complaints about selection biases and the representativeness of the data sets used to build the diagnostic, prognostic, and resource management–level prediction models in question tap into deeper concerns about how mismatches between data samples and target populations can lead to deleterious or discriminatory outcomes. It has long been recognized that insufficient cohort diversity and the under- or overrepresentativeness of data sets can lead AI/ML systems trained on this data to have biased and inequitable impacts on certain subpopulations (Barocas & Selbst, 2016; Calders & Zioblaite, 2013; Ferryman & Pitcan, 2018; Lehr & Ohm, 2017; O'Neil, 2017). Corresponding equity issues in data-driven approaches to medicine can, for example, arise in electronic health records (EHR) that fail sufficiently to include members of disadvantaged or marginalized groups who are unable to access the health care system (Arpey et al., 2017; Gianfrancesco et al., 2018) or in the sample selection biases that emerge when data availability is limited to well-resourced, digitally mature hospitals that disproportionately serve a particular racial or socioeconomic segment of a population to the exclusion of others.

Beyond data set inequities, health care relevant patterns of discrimination and bias arise throughout the AI/ML production workflow, from biased choices made in data preprocessing and feature engineering to the ways in which various model parameters are tuned over the course of model design and testing (Berk et al., 2017; d'Alessandro et al., 2017; Kamiran & Calders, 2012; Leslie, 2019a; Suresh & Guttag, 2020). Of particular concern are the potentials for discriminatory harm that surface at the level of problem formulation, namely, in the ways that data scientists and AI innovators define target variables and identify their measurable proxies (Passi & Barocas, 2019). Definition-setting determinations made by design teams and researchers can perpetuate and reinforce structural inequalities and structural injustices (Jugov &Ypi, 2019; Young, 1990, 2009, 2011) by virtue of biased assumptions that creep into the way solutions are devised and measurements molded. This form of discrimination can have an especially devastating effect in the field of health policy, as Obermeyer et al. (2019) demonstrated in their examination of how the label choice made for a commercial risk prediction tool in U.S. health care led to systemic discrimination against millions of Black patients, who tended to be far sicker than Whites at an equivalent risk score.[4]

If left unattended to—especially in view of the current design-time pressures placed on project teams for rapid responses and insights—these sociotechnical tendrils of algorithmic bias and discrimination may only further tighten their grip on AI-supported practices and outcomes. This ethical hazard is, in fact, made worse by the disproportionately harmful effects of the COVID-19 pandemic on disadvantaged and vulnerable communities that are already subject to significant health inequities as well as overly susceptible to catastrophic, disaster-related harms (Bolin & Kurtz, 2018; Fothergill & Peek, 2004; Klinenberg, 2015; Kristal et al., 2018; Van Bavel et al., 2020; Wang & Tang, 2020). Indeed, such coronavirus-linked harmful effects are creating a kind of vicious discriminatory double punch whereby existing biases that make inroads into the health care–related algorithmic tools and applications created to combat the illness may harm disadvantaged people more because of the high, safety-critical impact these technologies have on them simultaneously as these biases may harm more disadvantaged people due to the disproportionate damage being inflicted on them by the virus itself.

2.2. Pitfalls of COVID-19-Related Research II: Adverse Data Impact

A second issue raised by Wynants et al. (2020) has to do with data impact, namely, the need, in diagnostic, prognostic, and policy-level prediction models, for complete, consistent, accurately measured, relevant, and timely data that is of sufficient quantity to produce reliable out-of-sample generalization (Ehsani-Moghaddam et al., 2019; Heinrich et al., 2007; Wang et al., 1995). Wynants et al. highlight the fact that the studies that they appraise face the common hazard that small sample sizes (drawn from scarce and geographically limited patient populations) will lead to overfitting and compromised generalizability (Foster et al., 2014; Riley et al., 2020; Wynants et al., 2020). They note that "immediate sharing of well documented individual participant data from COVID-19 studies is needed for collaborative efforts to develop more rigorous prediction models and validate existing ones" (p. 7).

Extant issues with data quality and responsible data sharing in the health care domain will likely pose challenges here of which AI/ML researchers and innovators should take heed. Hindrances to the access and availability of sufficiently high-quality data in the context of a global public health emergency present difficulties that augment the effects of the widespread tendencies to health data silo-ing that have generated a motley of nonintegrated data formats (Shortliffe & Sepúlveda, 2018) and a wide variability in data quality and integrity (He et al., 2019; Hersh et al., 2013; Kruse et al., 2016; Verheij et al., 2018). Prevailing gaps in digital maturity across hospitals, regions, and countries may also act as roadblocks to accessing data of sufficient quality and quantity to pick up generalizable and transportable signals from target populations. The general lack of preparedness to mobilize digital information on the second front has already been evidenced in the scramble to rapidly hand-code COVID-19 symptom checker chatbots in lieu of training data accurate enough to pursue more sophisticated AI/ML methods (Kohler & Scharte, 2020).

Other data-mobilizing suggestions that have been made by AI/ML researchers confronting SARS-CoV-2-related clinical questions raise an equally vexing set of data quality and sharing concerns. Van der Schaar et al. (2020) have proposed to link EHRs with passive data from pervasive sensing and mobile technologies in order "to issue accurate predictions of risk and help uncover the social structures through which systemic risks manifest and spread" (van der Schaar et al., 2020, p. 2). While forward-looking, these proposals face obstacles in terms of data set representativeness and well-established uncertainties in the quality of unstructured big data (Bailly et al., 2018; Cahan et al., 2019; Kruse et al., 2016; Miotto et al., 2018). They also introduce long-concerning ethical risks related to informational privacy, deidentification, and informed consent—specifically, as these principles relate to the collection, linking, and use of passive data exhaust containing sensitive and potentially deanonymizing information (De Montjoye, 2015; Golle, 2006; Klasnja et al., 2009; Maher et al., 2019; Ohm, 2010; Smith et al. 2016; Sweeney, 2000).

2.3. Pitfalls of COVID-19-Related Research III: Lack of Process Transparency and Model Interpretability

A final set of issues raised by Wynants et al. (2020) can be grouped, by family resemblance, into the category of transparency. In the review, the authors emphasize the prevalence in the appraised studies both of the low quality and opaqueness of research methods and recording practices and of the opaqueness and lack of interpretability of the predictive models themselves. The first of these problems can be classified as insufficient process transparency, the second, insufficient outcome transparency (Leslie, 2019a). To take the former first, having transparent organizational and research practices as well as well-reported documentation of them becomes all-the-more vital in the context of the COVID-19 pandemic inasmuch as normal protocols that govern patient consent and privacy may be suspended, amended, or compromised. The absence of explicit sanction places a higher burden of transparency and accountability on researchers, who must ensure that their research practices are worthy of justified public confidence and trust. Even in recent, noncrisis times, concerns about a lack of this kind of process transparency, at the levels of both organisational conduct and research practice (van der Aalst et al., 2017), have prompted demands for better approaches to operationalizing answerability and auditability in health care–related AI/ML innovation, so that the public can be reassured that their health data are safely and responsibly being used for advancing patient and community wellbeing (Habli et al., 2020; Hays et al., 2015; Spencer et al., 2016; Stockdale et. al., 2019).

At a more basic level, anxieties about process transparency have already had a direct bearing on the adoption, application, and effectiveness of data-driven decision support in a broad range of clinical environments. Poor methodological conduct and reporting, across many different areas of research in clinical prediction modeling (Collins et al., 2013; Damen et al. 2016; Mallet et al., 2010; Siontis et al., 2015), has led to limitations in the perceived reliability and applicability of such studies in decision-support settings (Bouwmeester et al., 2012). Poor reporting and unclear methodological conduct function as a stumbling block for the reproducibility of results, and these are then often met with justified trepidations by clinicians. Without externally validated research that may be corroborated through replicated experimental methods and that supports generalizability to out-of-sample instances (Altman et al. 2009; Moons et al., 2012), clinical uptake will be significantly limited (Collins et al., 2014; Khalifa et al., 2019; Vollmer et al., 2020).

Similar degrees of reasonable mistrust have been generated among clinicians and patients due to a lack of transparency in the innerworkings and underlying rationale of the decision-assistance models themselves. As supports for evidence-based reasoning in medical and public health practices, diagnostic, prognostic, and policy-level prediction models bear the burden of having to be optimally intelligible, understandable, and accessible to clinical users and affected individuals (Academy of Medical Royal Colleges [AMRC], 2019; Doshi-Velez & Kim, 2017; Gilvary et al., 2019; Information Commissioner's Office & The Alan Turing Institute, 2020; Jia et al., 2020; Miller, 2019; Nauck & Kruse, 1999; Rudin, 2019; Vellido et al, 2012; Vellido 2019; Wainberg et al., 2018). The optimization of model interpretability enables data scientists to build explanatory bridges to clinicians (Lakkaraju, 2016), who can then draw upon a given model's processing results to justify evidence-driven clinical decision-making (Shortliffe & Sepúlveda, 2018; Tonekaboni et al., 2019). Such bridges allow for gains in the objectivity and robustness of clinical judgment by making possible the detection of a greater range of patterns drawn from the vast complexity of underlying data distributions accessible to practitioners (Morley et al., 2019). Moreover, a high degree of interpretability allows data scientists and end users to better understand why things go wrong with a model when they do; as such, it can help them to continually evaluate a model's limitations while scoping future improvements. Bridges between data science and clinical practice also allow clinicians to make sense of (and hence, to make better use of) inferences and insights derived from trained models in the contexts of their application domains

(Holzinger et al., 2017).[5] Some have even claimed that medical ethics require interpretable AI systems insofar as the doctors who use them to support their care of patients must be able to provide meaningful information about the logic behind the treatments chosen and applied (Vayena et al., 2018).

Regardless of the wide acceptance of these desiderata of AI/ML interpretability in medicine and public health, an unresolved tension remains within the concept of outcome transparency—one with significant ramifications for an innovation environment influenced by the exigencies of the pandemic response. This has to do with the oft discussed trade-off between performance (predictive accuracy or other metrics) and interpretability (Ahmad et al., 2018; Bologna & Hayashi, 2017; Breiman, 2001; Caruana et al., 2015; Freitas, 2013; Gunning, 2017; He et al., 2019; Holzinger et al., 2019; Selbst & Barocas, 2018). The conventional view suggests that the deployment of complex AI/ML model classes (like deep-learning or ensemble methods) leads, in general, to a boost in model performance in comparison to simpler techniques (such as regression- or rule-based methods), but only at the expense of interpretability. Thinking in the context of high-stakes decision-making, Cynthia Rudin has recently characterized the idea that this trade-off is necessary as a “myth” (Rudin, 2019, p. 2). She argues that in domains like medicine—where much of the clinical data are organically representable as meaningful features and well-structured—interpretable algorithms have roughly equivalent performance as more opaque techniques. At the same time, the native understandability of such algorithms eliminates the need for surrogate, post hoc explanatory mechanisms that tend to have low fidelity to the slippery nonlinearity of their black box counterparts.[6]

Rudin’s steer is away from an explainability culture (that begins by reaching for the black box and then tries to find simpler auxiliary models to elucidate it) and toward an interpretability culture that starts with attempts to produce interpretable models through solid knowledge discovery and careful model iteration.[7] This prioritization of outcome transparency is consistent with positions held by clinicians who see AI/ML decision-support systems as bolstering evidence-based medical practice by widening and enriching the informational background for the exercise of human judgment (Shortliffe & Sepúlveda, 2018; Tonekaboni et al., 2019). However, others have pointed out that the utility of high-performance, high-interpretability models in clinical environments has yet to be demonstrated due to the infrequency of their application (Ahmad et al., 2018). More significantly, while such models are clearly preferable when mining low-dimensional, structured data, some of the most medically consequential contributions of AI/ML systems have been based in the processing of complex, high-dimensional data by black box models (for instance in radiomics and medical imaging). Responding to this, Rudin has prospectively suggested that interpretation-aware methods such as in-building prototyping facilities can be integrated into even complex artificial neural nets (Chen et al., 2019; Li et al., 2018; Rudin, 2019), and, indeed, others have proposed that attention-based explainers be incorporated by design into model architectures of this sort (Choi et al., 2017; Park et al., 2017; Xu et al., 2018).[8]

The urgency of delivering rapid research responses to the COVID-19 pandemic puts a new kind of pressure on these emerging approaches to making complex, opaque models fit-for-purpose in supporting safety-critical decision-making. The development of new interpretability methods in clinical environments is likely to be put on the back burner, resulting in continued dependence on existing methodologies of explainability (for example, isolating the regions of interest in clinical imaging flagged by saliency maps or gradient-class activation maps). As Wynants et al. demonstrate, however, even the essential process of properly annotating medical images during deep-learning system design is not always well-executed in hasty research milieus (Wynants et al., 2020, p. 8), and time pressures placed on clinicians will challenge their capacities to thoroughly decipher and weigh up auxiliary explanation offerings. The outcomes of other potential applications of opaque model classes to unstructured, heterogenous data (or to combinations of this kind of data, say, free-text clinical notes, with EHRs) present explainability hurdles of their own. For instance, these may be explained by existing surrogate explanatory methods (like LIME or SHAP) that have been shown to have a spotty track record in generating accurate, reliable, and faithful accounts of the determinant features driving black box predictions (Alvarez-Melis & Jaakkola, 2018; Leslie, 2019a; Mittelstadt et al., 2019; Molnar, 2020). We should note, additionally, that all of these options for supplementary, post hoc explanation support do not yet address more fundamental concerns that, even if partially explainable, some opaque models may still bury error-inducing faults or patterns of discrimination deep within their architectures that may manifest in unpredictable, unsafe, or inequitable processing outcomes.

Although this enumeration of the difficulties faced by data scientists and AI/ML innovators is nothing new, a sense of urgency to confront them is. Taken together, the undercurrents of algorithmic bias, adverse data impact, and deficient process and outcome transparency are deep-rooted but open problems in data science that are presently made all-the-more challenging by the unprecedented pressures to tackle the pandemic. But these are problems with actionable solutions whose collective realization or evasion will be the historical axis that determines whether data science will be able to fulfil its massive potential to make a difference in the global fight against the virus. To set down a path toward this realization, data scientists will have to draw heavily upon the available moral-practical resources, existing knowledge, and sociotechnical self-understanding provided by current thinking in applied AI/ML ethics, social scientific approaches to data-driven technologies, and responsible research and innovation.

3. Digital Contact Tracing, Solutionist Lure or Public Health Tool?

Before moving on to exploring the proper direction that responsible AI research and innovation should take, however, we would do well to investigate a controversial set of pandemic-related data-driven technologies. As is widely known, data-driven applications are being developed to speed up contact tracing and manage contagion through targeted health surveillance and individual tracking, as well as to enable personalised approaches to societal reintegration as social distancing measures are eased. These kinds of applications are triggering a complex set of ethical hazards that are only exacerbating the mounting challenges to autonomy, privacy, and public trust already faced globally by citizens caught in the crucible of a ubiquitously networked, big data society. Nevertheless, these kinds of human monitoring and tracking applications may prove crucial for managing the rapid asymptomatic and presymptomatic transmission of the disease and for mitigating some of its more punishing social and economic consequences (Ferreti et al., 2020).

3.1. The First Wave of Digital Health Surveillance in Asia

A first wave of such interventions, taking place in the Asian countries first struck by the virus, has largely been characterized by a combination of the macroscale exercise of social control and the centralized consolidation of personal and mobile phone tracking data. In China, technologists have built a noncompulsory but ‘use-to-move’ AI application that integrates users’ personal, health, travel, and location data with public health information about SARS-CoV-2 cases to produce individualized risk scores (stratified into ‘health code’ levels of green, amber, and red). These determine who can access public spaces, shops, and public transport and who must be quarantined. The app, run through the prominent Alipay and WeChat platforms, is employed to monitor the movements of each of its roughly one billion users to ensure compliance and to keep continuous track of contacts (Calvo et al., 2020; Davidson, 2020; Mozur et al., 2020). Some reports out of China have been troubling. Not only may an AI-generated health code instantaneously turn from green to red without reason or explanation —as the algorithm behind the system is entirely opaque and made inaccessible to the public (Mozur et al., 2020)—a citizen caught traveling with a red code can be marked down in the country’s social credit system to devastating personal and professional consequence (Zhang et al., 2020).

In Taiwan, a strategy of data integration similar to China’s has been deployed that links the country’s national health insurance database with its immigration and customs database to assist health care workers in identifying probable cases of COVID-19 infection (Wang et al., 2020). Taiwan has also used AI to monitor travelers, who are assigned risk scores based upon the origin and history of their travels and subsequently trailed through their mobile phones to confirm fulfilment of quarantine restrictions (Lee, 2020). Such a digital monitoring method can be heavy-handed; if a phone possessed by a quarantined user dies or is turned off, a visit from the police will soon follow (Lee, 2020). A different tack has been taken in South Korea, where the government is mining vast troves of CCTV, financial, and phone-tracking data to reconstruct and publicize exhaustive—and potentially identifiable—logs of the movements and personal details of people who have tested positive for the virus (Zastrow, 2020). As has been noted by Nanni et al. (2020), such a method has marshalled data value to positive public health effect, while blatantly sacrificing patient privacy.

In contrast to South Korea, Singapore has taken a more consent-based and privacy-aware approach. It has implemented a Bluetooth-based proximity tracing system called TraceTogether—an opt-in decision-support application that helps public health officials to track down and communicate with the at-risk contacts of infected users (Bay et al., 2020). The TraceTogether app minimizes data collection by utilizing encrypted tokens that are exchanged between proximate users and then stored locally on their respective phones. Each user's nonpersonally identifiable tokens or "TempIDs" are issued by and stored on the server of the health authority, which also maintains a database of users' identifiable phone numbers. When individuals who have the app become ill with COVID-19, they are compelled by law to share their token exchange history with health officials, who are then able to use the central server to decrypt the tokens and compile a list of potentially infected users to contact (Bay et al., 2020; Cho et al., 2020).

3.2. The Coming Second Wave and the Prioritization of Privacy

This Singaporean model has set the scene for the second, privacy-sensitive wave of data-driven surveillance and tracking technologies that has begun to form in Europe, the United States, and beyond. Here, the direction of innovation has largely been steered by concerns about intrusive data collection, use, and repurposing by centralized governmental or commercial infrastructures. Such anxieties have shaped debates around perceived trade-offs between priorities of privacy, individual liberties, and data protection, on the one hand, and those of more collectively oriented values such as protecting public health and community wellbeing on the other. They have created an atmosphere of widespread apprehension that has led researchers and app developers to focus on finding technical solutions to the problem of optimizing privacy preservation while securing effective digital surveillance mechanisms.

Setting aside, for now, the ethical question of whether or not such a single-minded concentration on app-driven, technological solutions is justifiable given the plenitude of other relevant sociotechnical and practical factors at play (cf. O'Neil, 2020), two technical approaches have, so far, dominated the dash toward the development of privacy-preserving contact-tracing technologies: GPS-based methods of co-localization tracing (Berke et al., 2020; Ferreti et al., 2020; Fitzsimons et al., 2020; Raskar et al., 2020 [though this presents hybrid features]; Reichert et al., 2020) and Bluetooth-based methods of proximity tracing (Bell et al., 2020; Brack et al., 2020; Canetti et al., 2020; Chan et al., 2020; Cho et al., 2020; CoEpi, 2020; COVID Watch, 2020; Hekmati et al., 2020; PEPP-PT, 2020; TCN Coalition, 2020; Troncoso, 2020). In the former, the GPS location histories of diagnosed carriers are deidentified and encrypted before they are shared with a backend server that allows for other users' apps to check whether or not they have crossed paths with the infected individual (Berke et al., 2020; Raskar, 2020).

Proponents of this method argue that, despite some limitation of precision in determining collocation, its continuity with existing in-phone GPS-tracking facilities will streamline the ease of its adoption, allowing for the magnitude of uptake necessary to lower COVID-19's R_0, its reproduction number, below 1 (some estimate 3/5 of a total population) and to consequently achieve herd immunity (Ferreti et al., 2020). Berke et al. maintain that the app's technology can be "integrated into partnering applications that already collect user location histories, such as Google Maps":

> These partner applications can then ask the user for the extra permissions and content for this system's use case. There are many such applications that already collect user location histories in the background. They often use this information to serve the user more relevant content and improve the user experience. However, this data collection more often serves private profit. Now, in the face of the COVID-19 pandemic, is the time for industry and researchers to come together and for the ubiquitous collection of location data to serve the public good. (Berke et al., 2020, p. 11)

Among researchers and developers, however, there is increasing skepticism regarding the "significant privacy trade-offs" (COVID Watch, 2020) likely required in order for GPS-based methods of digital contact tracing to be functional. They have also flagged the correspondingly high computational burden placed on existing platforms by the cryptographic techniques needed to mitigate some of these issues (Bell et al., 2020; Chan et al., 2020). Another point of contention has been the accuracy limitations of location-centric methods—for instance, their inability to provide fine-grained recordings of interpersonal proximity and their lack of accurate functionality in certain buildings, subways, and multilevel dwellings. Such limitations call into question the capacity of GPS-based apps to measure human-to-human contacts with the degree of precision necessary to reflect medically defined specifications of disease exposure. These shortcomings have thus been taken to signal the advantages of Bluetooth-based techniques of proximity tracing (Canetti, 2020; Chan et al., 2020).

Unsurprisingly, Bluetooth-based tracing methods have now become the most likely data-driven technology to be applied to the health-surveillance dimension of the current coronavirus pandemic in Europe and the United States—a likelihood that has dramatically increased in light of an Apple/Google joint initiative to build a proximity tracing API for their three billion active mobile devices (Apple, 2020; Google, 2020). While human-in-the-loop proximity tracing technologies, like Singapore's TraceTogether, also use exchange and local storage of encrypted contact event tokens, many Anglo-European and American researchers are seeking to fully automate the Bluetooth-based contact-tracing process so that all reliance on 'trusted third parties' and data-consolidating central servers can be eliminated (Canetti et al., 2020; Chan et al., 2020; TCN Coalition, 2020; Troncoso, 2020).

In decentralized applications, users remain nonidentifiable to each other from beginning to end of any contact-tracing process. Data shared with central servers is minimized, and all contact detection, infection discovery, and risk computation are locally initiated and processed. When infection carriers are diagnosed, they receive health authority authorizations that then enable their anonymized contact histories to be uploaded onto a backend server. Meanwhile, the apps of other users periodically query this server to see if there are any contact matches. In the event that there are, each smartphone calculates a risk score and decides whether or not notification of its user is appropriate (esp. Troncoso, 2020). By taking humans out of the loop through this type of comprehensively automated decentralization—a strategy whereby algorithmic models independently determine individual risk without the provision to users of any specific information backing or justifying the decision—the threats of adversary infrastructures that may violate privacy rights and infringe on data protections (whether they be governmental or commercial) are believed to be mitigated.

3.3. Public Health Priorities of Past Digital Health Monitoring Interventions

Although the COVID-19-era emphasis on privacy and data protection has offered an important counterforce to the globally consequential threat of mass surveillance, questions remain as to whether this narrow focus on building airtight technological solutions has diverted attention away from some of the more salient underlying motivations and complexities that surround the introduction of digital contact-tracing innovations during public health crises. Notably, the use of data-driven technologies to provide this kind of assistance was originally framed under the rubric of advancing digitally supported mobile health (mHealth) in the end of safeguarding community wellbeing (Danqua et al. 2019; Ha et al., 2016; Reddy et al., 2015; Mendoza et al., 2014; Sacks et al., 2015; WHO, 2011; Zhenwei Qiang et al., 2012). While several earlier approaches did prioritize privacy-preserving methods of digital contact tracing (Altuwaiyan, 2018; Prasad & Kotz, 2017; Shahabi 2015), the primary aim in pilot implementation studies and research interventions in this area was to optimize technological support for medical responses to ongoing epidemics. Thus, in Sacks et al. (2015), a smartphone-based mHealth tool was introduced to assist public health officials with Ebola surveillance and contact tracing in Guinea during the 2013–2015 epidemic. Though the tool faced significant adoption challenges, it “offered potential to improve data access and quality to support evidence-based decision-making for the Ebola response” (Sacks et al., 2015, p. 646).

Similar mHealth interventions were made by Ha et al. (2016) to assist with tuberculosis contact tracing in Botswana in 2012 and by Danqua et al. (2019) during a 2015 Ebola outbreak in the Port Loko district of Sierra Leone. In the latter study, an Ebola contact-tracing app was successfully deployed to streamline and support communication between contact tracers and the public health coordinators. Responding to the 2013 dengue outbreak in Fiji, Reddy et al. (2015) introduced a GPS-based mHealth phone-tracking tool that both helped public health officials to pinpoint infected areas and patients to become better informed about the symptoms of the disease and its treatments. The app also encouraged community-led support of public health efforts through cooperative involvement in reporting and identifying hotspots. Its creator concluded that the tool was “not going to replace physicians, however, it will greatly assist them in making their work easier in controlling disease outbreaks” (Reddy et al., 2015, p. 12).

3.4. The Solutionist Lures of Automation All-the-Way-Down

A different, and potentially counterproductive, approach to digital contact tracing has developed in the context of the present coronavirus pandemic as normative emphasis has shifted from a stress on supporting medical response effectiveness to an emphasis on the extent to which the privacy-upholding expansion of automation can appease misgivings about state adversaries, mass surveillance, and function creep. The politics of public distrust may be purging the priority of the ‘public’ from the province of public health per se and providing an impetus to automation all-the-way down. While a data-minimizing, privacy-preserving perspective on digital contact tracing is vital to its justifiability and societal acceptance, privacy-first apps that side-step the third-party, human-in-the-loop involvement of trusted contact tracers in investigation and risk determination should give us pause.

There are several reasons for this hesitation. First, a reliance on fully automated contact-tracing methods for data collection and evaluation as well as for subsequent risk determination may betoken overconfidence in that data’s accuracy, precision, and integrity and, consequently, in the reliability of the system that processes them. Common weaknesses in the integrity and quality of sensor data collected by digital devices (Ienca & Vayena, 2020) limit the likelihood that Bluetooth-based contact-tracing technologies will be able to meet the high bar of functional requirements set by system designers themselves. In particular, barriers to measurement accuracy stemming from Bluetooth signals that fail to take account of glass windows, room dividers, product-lined shelves separating supermarket aisles, thin walls, mask-wearers, and so on (Fussell & Knight, 2020) raise questions as to whether this kind of contact-tracing app can meet stated requirements such as “precision” (i.e., that “reported contact events must reflect actual physical proximity”) and “integrity” (i.e., that “contact events are authentic”) (Troncoso, 2020, p. 3).

To cut human contact tracers out of a public health process that is then bound to overrely on fully automated tracing technologies is to preclude the application of common sense, context-awareness, and skilled judgment in remediating the data quality and integrity issues that will inevitably arise and in authenticating the veracity of data-processing results. This difficulty is amplified in epidemiological settings, where the outcome-determinative gradients of encounters between infection carriers and at-risk individuals (close, casual, or transient) are often highly dependent on the contextual nuances of such factors as location and environment—nuances simply unavailable to the rigid algorithmic models behind contact-tracing apps' risk calculations (Bay et al., 2020). For example, "short-duration encounters in enclosed spaces without fresh ventilation often constitute close contact, even if encounter proximity and duration do not meet algorithmic thresholds" (Bay et al., 2020, p. 6). Without the availability and use of common sense and human discernment, vital distinctions that might help public authorities avoid both false positives and false negatives will be lost. Such a judgment gap in the implementation of fully automated contact-tracing systems suggests that the inferential brittleness of these apps may lead to ineffective or even deleterious 'garbage-in-garbage-out' results. This would likely produce a generalized unease about adopting such technologies given the significant possibility that they will produce erroneous outcomes at the cost of either personal health or freedom of movement.

Furthermore, such a dislodging of human judgment raises the graver concern that taking humans out of the loop may, in fact, contribute to the deterioration of social trust in the public health authorities charged with handling public health emergencies. Though crucial to take into consideration, the adversary assumption that eschews any trusted third party and motivates the comprehensive decentralization of digital contact-tracing technologies is, perhaps, insufficiently attentive to the delicate role played by reciprocal relations of social trust and interpersonal responsibility in establishing and sustaining the fabric of shared confidence and mutual reliance that undergirds effective and community-involved public health responses to public health crises.

The reason for this runs deep. Taken together, trust and responsibility have formed an implicit normative pillar of social order in the modern era. When individuals behave and act in ways that affect one another for better or worse, contemporary society binds them to the justifiability of their actions based upon reasonable expectations that, as rational agents, they will exercise good judgment in pursuing their objectives in ways that do not harm those around them and that are made accountable in virtue of such a "generalised expectancy" (Rotter, 1967). Securing such a nexus between the responsibility of each and the trust of others involves establishing a bedrock of situation-independent behavioral expectations between rational agents whereby mutually accountable performances can be universally assumed (Bauer & Freitag, 2018). Such a stable starting point for a free but orderly social coexistence has been variously called "basic trust" (Erikson, 1959) and "generalised trust" (Uslaner, 2002).

The problem with the complete automation of contact tracing is that it would do away with the architectonics of reasonable expectation that serve as an underpinning of generalised trust in the domain of public health. When crucial health decisions, such as a quarantine determination after an assessment of potential exposure, are taken out of the custody of responsible public health professionals, the kinds of reasonable expectations that anchor public trust (both in the institution and in the process) are likewise removed from the picture. Instead, a smartphone vibrates with an impersonal alert that perforce remains inexplicable to the potentially infected decision subject in its details and rationale. No reasonable expectations are involved inasmuch as one cannot, in effect, have these if there are no reasons on offer in the first place. And, when relations of reciprocal responsibility are consequently replaced by a *vox ex machina*, the unquestionable force of preemptory calculation leaves blind obedience to the algorithmic result as the only practicable option. Counterintuitively, this upshot of decentralization may have a kind of panoptical effect where a bloodless notification of infection risk on a mobile app punctuates a continuous dynamic of depersonalizing digital surveillance. At the negative extreme, this would mean that self-reinforcing mechanisms of social distrust end up optimizing privacy at the expense of creating conditions of deteriorated autonomy, social connectedness, and solidarity.

By contrast, in environments where research and innovation practices are organized around optimizing medical responses to public health emergencies and thus more directly oriented to the priority of societal wellbeing, digital contact-tracing apps are seen as supporting evidence-based but compassion-driven human decision-making. For example, the creators of Singapore's TraceTogether app have stressed the importance of a humane and human-centered approach: "Contact tracing involves an intensive sequence of difficult and anxiety-laden conversations, and it is the role of a contact tracer to explain how a close contact might have been exposed—while respecting patient privacy—and provide assurance and guidance on next steps" (Bay et al., 2020, p. 7). Here, the second-front design and deployment of a decision-supportive contact-tracing technologies is understood to enable frontline contact tracers to "incorporate multiple sources of information, demonstrate sensitivity in their conversations with [citizens] who have had probable exposure to SARS-CoV-2, and help to minimize unnecessary anxiety and unproductive panic" (Bay et al., 2020, p. 7).

3.5. Privacy, Public Health, and Power

The contrast between the emerging privacy-first approach taken by proponents of fully automating digital contact-tracing apps and the more public health-centered perspective instantiated in the Singaporean TraceTogether-supported method returns us to the debate around the seemingly unavoidable trade-offs between privacy and individual liberties, on one side, and community wellbeing and societal benefit, on the other. The difficulties faced at the extremes of the debate—at one end, the potential for radically centralized forms of health surveillance to lay waste to fundamental rights and freedoms, and at the other, the potential for radically decentralized forms to reinforce social distrust and to harm individual autonomy and interpersonal solidarity—should perhaps draw our attention to an additional factor that must also be considered. This is the issue of power as it relates to the use of data-driven technologies: What is the legitimate scope of the exercise of power during times of crisis and emergency? What are the real possibilities for its abuse or misuse, both by governments and by private companies, in regard to digital contact tracing and surveillance? What are the short- and long-term consequences of its impingement upon the digital organs that now sustain so much of our networked and connected private lives?

These questions highlight how the problem of power inexorably shades any consideration of the ethical challenges presented by digital contact tracing or beneficent health surveillance in contemporary big data society. Though the legitimacy of these sorts of data-driven technological interventions largely hinges on building reason-based public confidence in the appropriateness and justifiability of their employment, we do not, at present, live in a culture of public trust, when it comes to data collection, sharing, and use. The longstanding monetization of personal data by Big Tech companies has left members of society reasonably skeptical about how their data is being extracted and appropriated (Fourcade & Healy, 2013, 2017; Fuchs, 2010; Sadowski, 2019; Srnicek, 2017; Zuboff, 2015, 2019). After years of having algorithmically personalized services reach into their private lives to curate their tastes, nudge their behaviors, and steer their consumption, data subjects are sensibly on guard. Add to this the frightening but all-too-common instances, in many parts of the world, of intrusive governmental use of algorithmic targeting and manipulation for purposes of social control (Creemers, 2018; Roberts et al., 2019; Wright, 2019a 2019b), and it becomes easy to understand trepidation about the deployment of digital monitoring, tracking, and surveillance (Russell, 2019).

In the context of the second-front fight against COVID-19, attention to questions about power should key us in to the central importance of instituting regimes of responsible AI innovation in order to establish, and convince citizens about, the ethical justifiability, trustworthiness, and public benefit of such interventions. If data are to be legitimately marshaled through digital contact tracing, and health surveillance is to serve the purposes of community wellbeing, such innovations will have to be proportionate, socially licensed, and democratically governed. Normative AI regimes should ensure that research and innovation processes are reflective in anticipating ethical and societal impacts, that they are informed, from the start, by inclusive and collaborative deliberations on the balancing of potentially conflicting values, and that they are context-aware, domain-knowledgeable, and codesigned with the individuals and communities they affect. Such digital innovations will thus have to be explicitly values-driven, consent-based, and shaped by open public dialogue. Their processes of design and deployment will require transparency, continuous public oversight, rigorous pilot testing, reflective integration into wider public health strategies, and well-defined limitations. Developed responsibly, such technologies will have to be reasonably privacy preserving, compliant with human rights and responsible data management protocols, and subject to sunset and retirement provisions, which set clear and predefined constraints on their application to the present exceptional circumstances of the pandemic.

4. Five Steps Toward Responsible AI Innovation

A focus on responsible AI innovation, in the context of digital contact-tracing and tracking apps, shows that it is essential not to fall prey to a tempting but false choice. This is between a sense that, in order to use these technologies, we must relinquish our fundamental rights and freedoms to the strengthening powers of the surveillance state and a sense that we must protect our privacy and individual liberties at the cost of pressing the full capacities of our data-driven technologies into the service of the public good. Both of these all-or-nothing alternatives fail to discern the potential of socially licensed innovation to function as a progressive counterforce to the excessive exercise of power. The potential rise of digital autocracies and AI-enabled totalitarian regimes, the abusive data grabs of state adversaries and profit-oriented commercial entities, the preemptive manipulation of human behavior by platformed algorithmic infrastructures, these are real problems. But they are problems that modern free societies must combat by harnessing the democratic energies of open communication, public engagement, and collaborative value articulation. Drawing upon and strengthening the reflective, inclusive, and participatory character of practices of responsible innovation is, in fact, one of humanity's most effective instruments to accomplish this crucial task.

Even in the case of digital contact tracing and individual tracking, the cooperative steering and democratic governance of technology should, in this respect, be seen as a potential source of citizen empowerment and community-involving public health support rather than a fast track to despotic surveillance. In our time of pandemic, as leaders of nation-states take hold of extensive emergency powers, the deterioration of the rule of law, the possibility of the abuse of unchecked authority, and the potential for 'surveillance creep' are hazards that merit sustained critical attention (Calvo et al.,2020; French & Monahan, 2020). But, grim as they may be, these are political possibilities rather than societal inevitabilities, and they must be met head-on by innovators, researchers, and citizens alike with the humane, communicative, and rational spirit of modern science. As nearly five centuries of the modern scientific method have shown, the open, dialogical, and consensus-based character of innovation processes are both a practical and epistemological necessity—a condition of possibility of the success of science itself (see Appendix).

Considering all this, a starting point in practices of responsible innovation should be embraced as a first priority for those in the data science and AI/ML community, who are doing battle on the second front of the global struggle against COVID-19. Fortunately, the scientific community does not have to fly blind in figuring out how to meet these exigencies of ethical research and discovery. For almost half a century, concerted efforts to flesh out responsible ways of pursuing the design and use of increasingly powerful technological tools have been made in areas ranging from bioethics (Beauchamp & Childress 2001; Department of Health, Education, and Welfare, 1974; Kuhse & Singer, 2009) and responsible research and innovation (RRI) (Hellström, 2003; Owen et al., 2012; Von Schomberg, 2011, 2013, 2019) to applied ethics (May & Delston, 2016; Singer, 1979, 1986), science and technology studies (STS) (Jasanoff, 2012, 2016; Sengers et al., 2005; Star, 1999), and, more recently, digital and AI/ML ethics (Engineering & Physical Sciences Research Council [EPSRC], 2011; Floridi, 2010; Floridi & Cowls, 2019; Jobin, 2019; Leslie, 2019a; Zeng et al., 2019).

We might do well, then, to turn to this body of research as a way to start upon a much longer journey toward creating a culture of responsible innovation in the data science and AI community. For this reason, I want to move now to proposing five steps that should be taken in order to responsibly bring the insights of data science and the tools of AI/ML to bear on the wide range of biomedical, epidemiological, and socioeconomic problems raised by the coronavirus pandemic. When incorporated into research and innovation processes from the start, these best practices will not only enhance the quality of research and discovery without adding undue burdens, they will improve the quality of outcomes and results. To put it simply, *responsible* data science is *good* data science—data science *with* and *for* society and worthy of public trust.

Step I: Open Science and Share Data Responsibly

Open science and *open research* build public trust through reproducibility, replicability, transparency, and research integrity (European Commission, 2014; McNutt, 2014; National Academies of Sciences, Engineering, and Medicine, 2018, 2019; Nosek et al., 2015; The Turing Way, 2019). The cooperative and barrierless pursuit of scientific discovery accelerates innovation, streamlines knowledge creation, fosters discovery through unbounded communication, and increases the rigor of results through inclusive assessment and peer review (Fecher & Friesike, 2013). Opening models and research procedures to expert assessment, oversight, and critique allows for rapid error and gap identification and catalyses the improvement of results. Moreover, reproducible and replicable research that is made accessible to all helps create confidence across society in the validity of scientific work.

The global reach of open research to an unbounded scientific community is especially important in the battle against COVID-19. The coronavirus pandemic is a species-level crisis, and so the scope and cooperative reach of the practices of scientific ingenuity that seek to redress it should also be global and inclusive. Managing the spread of the infection effectively will involve bolstering the knowledge as well as the control and mitigation strategies of every nation great and small.

A crucial constituent of this global effort is *responsible data sharing***.** While the first critical step in this direction is to *open up data* so that research can be reproduced and reused, data sets can be iteratively improved, and investments of time and research funding can feed forward to keep benefitting the public good (Borgman, 2015; Burgelman et al., 2019; Molloy, 2011; Piwowar et al., 2011; Tenopir et al., 2011; Whitlock, 2011), the concept of 'open data' itself must be bounded and qualified (Dove, 2015;

Jasanoff, 2006; Leonelli, 2019). Data sharing does not occur in a sociocultural, economic, or political vacuum but is rather situated amid an interconnected web of complex social practices, interests, norms, and obligations. This means that those who share data ought to practice critical awareness of the moral claims and rights of the individuals and communities whence the data came, of the real-world impacts of data sharing on those individuals and communities, and of the practical and sociotechnical barriers and enablers of equitable and inclusive research.

First and foremost, data sharers have a responsibility to serve the interests of wider society through the ethically piloted advancement of science, while simultaneously protecting the privacy and interests of affected data subjects. Accessible, high-quality, and well-archived data are the most critical ingredients in the progress of data scientific insights and AI/ML technologies, but responsibly opening data also involves privacy optimized, impact aware, and security-compliant data sharing. These two components can be seen as complementary: Properly managed accessibility and maximal data integrity allow for trusted data to be more freely circulated among an ever-widening circle of responsible researchers so that results can be replicated, and new, societally beneficial insights produced.[9] Responsible research that moves in this direction should refer to well-established protocols for responsible data management like those of the FAIR data principles (findable, accessible, interoperable and reusable data) (Wilkinson et al., 2016), trusted digital repositories (ISO 16363), Criteria for Trustworthy Digital Archives (DIN 31644), and the Data Archiving and Networked Services' CoreTrustSeal.

Data scientists and AI researchers who are tackling COVID-19 should also take heed of the higher demands for *data integrity* in safety-critical and highly regulated environments like health care. Data integrity, in this vein, can be understood as those dimensions of responsible data governance that safeguard trustworthiness across the entire data lifecycle from collection and correction through processing and retention. A useful framework for responsible end-to-end data governance is the "five safes" published by the United Kingdom's Office for National Statistics (Desai et al., 2016). The five safes aim to ensure that data is used for a morally and legally justifiable purpose and for the public benefit (safe projects), that researchers are well-trained and can be trusted to use the data appropriately (safe people), that the data is reliably deidentified (safe data), that access to the data is managed in a secure and situationally appropriate way (safe settings), and that research outputs are nondisclosive and do not provide opportunities for re-identification (safe outputs). Additionally, a high bar for standards of data integrity can be found in the "ALCOA plus" principles, which have been condoned and described in helpful guidance on data integrity (in the context of pharmaceuticals and medical devices) produced by the WHO and by the UK's Medicines and Healthcare products Regulatory Agency (MHRA, 2018; WHO, 2014).

The special responsibilities shouldered by researchers who are trying to apply responsible data-sharing practices in a global public health crisis has been broached in the WHO's 2015 consultation, *Developing global norms for sharing data and results during public health emergencies* (Modjarrad, 2016). Here, the WHO stresses that "timely and transparent prepublication sharing of data and results during public health emergencies must become the global norm" (WHO, 2015, intro.). Moreover, it affirms that opting in to the sharing of data and data analyses must be treated as a default practice and a moral obligation:

Every researcher that engages in generation of information related to a public health emergency or acute public health event with the potential to progress to an emergency has the fundamental moral obligation to share preliminary results once they are adequately quality controlled for release. The onus is on the researcher, and the funder supporting the work, to disseminate information through pre-publication mechanisms, unless publication can occur immediately using post-publication peer review processes. (WHO, 2015, para. 2)

Notwithstanding the WHO's endorsement of open research and responsible data sharing, authors of the background briefing prepared for the 2015 consultation identified several barriers to information sharing (Goldacre et al., 2015) that also figure in the context of the COVID-19 pandemic. These include issues related to information governance and data protection when ambiguities arise regarding informed consent and the confidentiality of potentially re-identifiable personal data, tensions between the need to share results rapidly and risks of inaccurate information doing harm in clinical environments, legacies of proprietary protectionism and the chilling effects of motivations to hoard data in the ends of academic publication priority, and delays in data sharing caused by the lengthy peer review processes involved in scientific journal publication.

Though many of these issues may be addressed through deliberate attitude change and the institution of governance regimes that ensure transparency and accountability, other barriers to responsible data sharing are rooted in more intractable social formations, such as underlying global inequalities and territorially and regionally variant political priorities that undermine federated, international approaches to addressing public health emergencies through open research. These are presenting scientists and innovators combating SARS-CoV-2 on the global plane with difficulties that are less immediately soluble but that should nevertheless be kept in view.

To take the issue of political priorities first, fears of outbreak-related reputational damage, migration and trade restrictions, widespread social stigma, damage to financial markets, and exposure of national security vulnerabilities may lead countries, political leaders, and state-controlled institutions to dissemble data and to clamp down on information dispersion. Varying instances of this occurred in the 2003 SARS CoV outbreak, in the 2009 H1N1 influenza pandemic, and in recent cholera outbreaks (Briand et al., 2011; Goldacre et al., 2015; Huang, 2004). Despite the explicit reporting and information-sharing provisions in the WHO's (2005) *International Health Regulations*, the high economic and geopolitical stakes of global public health emergencies can motivate political actors to engage in obstructive behaviors that prevent forthright and unhindered data dissemination.

Heeding these possibilities, data scientists and AI innovators must prioritize boots-on-the-ground communication with the researchers, clinicians, and domain experts who are directly involved in responding to and gathering data about the COVID-19 pandemic. More than that, innovators should bear in mind these political factors when they critically assess changes in the data landscapes as the current global public health crisis runs its course.

A second barrier to responsible data sharing, to which data scientist and AI/ML innovators should pay close attention, originates in long-standing dynamics of global inequality that may undermine reciprocal sharing between research collaborators from high-income countries (HICs) and those from low-/middle-income countries (LMICs). Given asymmetries in resources, infrastructure, and research capabilities, data sharing between LMICs and HICs, and the transnational opening of data, can lead to inequity and exploitation (Bezuidenhout et al., 2017; Leonelli, 2013; Shrum, 2005). For example, data originators from LMICs may put immense amounts of effort and time into developing useful data sets (and openly share them) only to have their countries excluded from the benefits of the costly treatments and vaccines produced by the researchers from HICs who have capitalized on such data (Goldacre et al., 2015).

Moreover, data originators from LMICs may generate valuable data sets that they are then unable to independently and expeditiously utilize for needed research, because they lack the aptitudes possessed by scientists from HICs who are the beneficiaries of arbitrary asymmetries in education, training, and research capacitation (Bull et al., 2015; Merson et al., 2015). This creates a two-fold architecture of inequity wherein the benefits of data production and sharing do not accrue to originating researchers and research subjects, and the scientists from LMICs are put in a position of relative disadvantage vis-à-vis those from HICs whose research efficacy and ability to more rapidly convert data into insights function, in fact, to undermine the efforts of their disadvantaged research partners (Bezuidenhout et al., 2017; Crane, 2011).

This challenge of misshapen reciprocity brings out a deeper issue pertaining to the framing of the desideratum of open data. While the challenge of overcoming the problem of global digital inequality in the era of data-driven innovation has often been approached under the rubric of traversing the 'digital divide' through more equitable provision of the resources needed to access information and communication technologies (ICTs), such a perspective neglects the enabling conditions of the globally diverse and disparately resourced *practices of innovation* that are needed to convert such technological resources into insights and applications. It is important, that is, to quarry beneath the issues of resource availability and allocation of ICTs, which have largely framed the impetus to opening data, and to concentrate as well on what Bezuidenhout et al. refer to as "the infrastructural, social, institutional, cultural, material and educational elements necessary to ensure the realization of openness" (Bezuidenhout et al., 2017, p. 465).

On this view, in redressing the barriers of inequality that hamper the responsible opening of data, emphasis must be placed on "the social and material conditions under which data can be made useable, and the multiplicity of conversion factors required for researchers to engage with data" (Bezuidenhout et al., 2017, p. 473). Equalizing know-how and capability is a requisite counterpart to equalizing access to resources, and both together are necessary preconditions of ethical data sharing. With this in mind, data scientists and AI/ML innovators engaging in international research collaborations should focus on forming substantively reciprocal partnerships where capacity-building and asymmetry-aware practices of cooperative innovation enable participatory parity and thus greater research equity.

Step II: CARE & Act Through Responsible Research and Innovation (RRI)

This demand for researchers to be responsive to the material and social preconditions of responsible innovation practices reminds us of the wider practical purview of RRI. An RRI perspective provides researchers and innovators with a vital awareness that all processes of scientific discovery and problem-solving possess sociotechnical aspects and ethical stakes. Rather than conceiving of research and innovation as independent from human values, RRI regards these activities as morally implicated social practices that are duly charged with a responsibility for critical self-reflection about the role that such values play in discovery, engineering, and design processes and in consideration of the real-world effects of the insights and technologies that these processes yield.

The RRI view of 'science with and for society' has been transformed into helpful general guidance in such interventions as EPSRC's 2013 AREA (Anticipate, Reflect, Engage, Act) framework and the 2014 Rome Declaration. These emphasize the importance of anticipating the societal risks and benefits of research and innovation through open and inclusive dialogue, of engaging with affected stakeholders as a means to co-creation at all stages of the design, development, and deployment of emerging technologies, and of ensuring transparent and accessible innovation processes, products, and outcomes (Owen, 2014; Owen et al., 2012).[10] The AREA framework is a handy tool to continuously sense check the social and ethical implications of innovation practices. Adding to this the priority of contextual considerations, we have the CARE & Act Framework:

> **Consider context**—think about the conditions and circumstances surrounding research and innovation; Focus on the practices, norms, and interests behind it. Take into account the specific domain in which it is situated and reflect on the concrete problems, attitudes, and expectations that derive from that domain;

> **Anticipate impacts**—describe and analyze the impacts, intended or not, that might arise. This does not seek to predict but rather to support an exploration of possible risks and implications that may otherwise remain uncovered and little discussed;

> **Reflect on purposes**—reflect on the goals of, motivations for, and potential implications of the research, and the associated uncertainties, areas of ignorance, assumptions, framings, questions, dilemmas and social transformations these may bring;

> **Engage inclusively**—open up such visions, impacts and questioning to broader deliberation, dialogue, engagement and debate in an inclusive way. Embrace peer review at all levels and welcome different views; and

> **Act responsibly**—use these processes to influence the direction and trajectory of the research and innovation process itself. Produce research that is both scientifically and ethically justifiable. (EPSRC, 2013, amended and expanded)

The CARE & Act Framework provides an actionable way to integrate anticipatory reflection and deliberation into research and innovation processes, while also emphasizing that an earlier stage-setting step must be taken to enable such an approach. Building this bridge from context to anticipation, reflection, and engagement is crucial. A solid understanding of innovation context is a precondition of effective anticipatory reflection inasmuch as it provides access to the key domain- and use-case-specific needs, desiderata, obligations, and expectations that frame preemptory considerations of the potential risks and impacts of any given research and innovation project. For instance, domain-situated knowledge of an AI system's operating environment will yield useful information about relevant industry standards and norms, organizational and public expectations, and outcome-influencing social factors and circumstantial exigencies. By taking contextual aspects like these into account, researchers and innovators will be better able to weigh up risks and impacts, to elicit the design and implementation requirements that address or mitigate them, and to take deliberate design-time actions to meet these requirements.[11]

There is one other component of RRI's capacity to build the bridge from context to anticipation, reflection, and engagement that is important to mention. Fruitful efforts to integrate the embodied, interactive, and pragmatic perspective of human–computer interaction (HCI) scholarship into RRI have helped to highlight the importance of contextual self-awareness and situational responsiveness in responsible innovation practices (Eden et al., 2013; Grimpe et al., 2014; Jirotka et al., 2017; Stahl & Coeckelbergh, 2016). In particular, reflexivity and anticipation are seen, from this standpoint, as concretely enacted amid the needs, opportunities, and problems of the particular communities of practice in which innovators and researchers are embedded (Grimpe et al., 2014). This means that contexts of innovation are animated for these innovators and researchers through their responsiveness to real-world challenges and to the continual demands of collaborative problem-solving. Such a de-idealized mode of "reflection-in-action" (Grimpe et al., 2014, p. 2972) consequently enables practices of RRI to stay warm-blooded and agile as scientists and innovators face the novel ethical difficulties posed by unforeseen problems and unknown unknowns.

To tackle COVID-19 responsibly, data science researchers and AI/ML innovators will have to marshal this agility and situational responsiveness as they cope with the innovation context of the present global health crisis. Helpful resources for gaining a general working understanding of this contextual dimension can be found in the WHO's *Guidance for Managing Ethical Issues in Infectious Disease Outbreaks* (2016), in the Council for International Organization for Medical Sciences' *International Ethical Guidelines for Health-Related Research Involving Humans* (2017), and in the Nuffield Council on Bioethics' *Research In Global Health Emergencies: Ethical Issues* (2020). Against the specific backdrop of data science and AI innovation, the following nonexhaustive list of contextual considerations may help orient anticipatory reflection within the frame of the coronavirus pandemic:

> **Magnified harmful effects on vulnerable and disadvantaged communities**—As we are already seeing in the devastating impact of the SARS-CoV-2 outbreak on communities of color and impoverished social groups, the pandemic is disproportionately affecting members of our society who are subject to structural legacies of disadvantage that put them at greater risk than others. When designing and developing innovation, researchers must take heed of these circumstances of vulnerability and disadvantage (MacIntyre & Travaglia, 2015). They must focus on protecting those who are most at risk and on ensuring that technological interventions purposefully yield societally equitable outcomes.

> **Disruption of public order and social, moral, political, and legal norms**—The governmental exercise of emergency powers and the urgency of producing swift and effectively scaled responses to public health crises can disrupt public order, societal norms, and the rule of law. This may occasion abrupt and wide-scale social changes, which subsequently have deleterious or regressive long-term consequences. For instance, if pursued without predefined limitations, the enforcement of censorship and surveillance measures to protect the public during an outbreak could shift norms of public acceptability and legal protections away from the safeguarding of civil liberties and fundamental rights and freedoms. Scientists and innovators should proceed with vigilance in analyzing the protracted effects of the innovations they produce.

> **Compromised consent and decision-making**—Public health crises put affected individuals and communities as well as frontline care providers under conditions of extreme duress, urgency, and distress (British Medical Association, 2020; Nuffield Council on Bioethics, 2020; WHO, 2016b). Those who are stricken with infection, or have sick family members, have to cope with uncertainty, suffering, fear, and powerlessness—all of which can compromise the processes of assessment, deliberation, and judgment that are required for the provision of informed consent. Likewise, in overwhelmed clinical environments, health care professionals are faced with constant demands to render critical decisions under conditions of incomplete information, extreme urgency, uncertainty, and disorder. Scientists and innovators must carefully take into account the distressed circumstances of those impacted by their research, and they must, where possible, prioritize ways of gaining informed and voluntary consent that accommodate these challenges, while also respecting the dignity of every person, recognizing their unique hardships, and taking into account the reasonable expectations of impacted individuals (consistent with Barocas & Nissenbaum, 2014; Nissenbaum, 2009). Likewise, innovators who are designing AI/ML decision-support systems for distressed clinical environments must take into consideration their distinctive implementation needs.

Step III: Adopt Ethical Principles to Create a Shared Vocabulary for Balancing and Prioritizing Conflicting Values

In our pluralistic and culturally diverse world, resolving ethical dilemmas is often dependent on building inclusive and well-informed consensus rather than appealing to higher authorities or to the say-so of tradition. This need for consensus-building is especially crucial in the context of AI/ML innovation, where circumstances often arise in which ethical values come into tension with each other. For instance, there may be situations (such as with digital contact tracing) in which the use of data-driven technologies may advance the public interest only at the cost of safeguarding certain dimensions of privacy and autonomy. Trade-offs, in cases like these, may be inevitable, but, regardless, the choices made between differing values should occur through a medium of equitable deliberation, mutual understanding, and inclusive and knowledgeable communication.

To this end, it is especially important to set up procedural mechanisms that enable reciprocally respectful, sincere, and open dialogue about ethical challenges. These mechanisms should help conversation participants speak a common language so that, when an innovation project's potential social and ethical impacts are being assessed and reassessed, diverging positions can be weighed and reasons from all affected voices can be heard, understood, and suitably considered. This can be accomplished by adopting common ethical principles from the outset to create a shared vocabulary for informed dialogue about balancing conflicting values.

There is, however, an obvious and important stumbling block that must be dealt with by this point of view. Amid the kaleidoscopic plurality of modern social life, no fixed or universally accepted list of ethical values could prereflectively provide such a common starting point. Researchers in AI/ML ethics have therefore had to take a more pragmatic and empirically driven position, in proposing basic values, that begins by considering the set of real-world problems posed by the use of the AI/ML and data-driven technologies themselves. These hazards include the potential loss of human agency and social connection in the wake of expanding automation, harmful outcomes that may result from the use of poor-quality data or poorly designed systems, and the possibility that entrenched societal dynamics of bias and discrimination will be perpetuated or even augmented by data-driven AI/ML technologies that tend to reinforce existing social and historical patterns.

In responding to such hazards, dozens of frameworks in AI/ML ethics have, over the past few years, more or less coalesced around a set of principles originating in both bioethics and human rights regimes (for example, Floridi & Clement-Jones, 2019; High Level Expert Group on AI, 2019; Institute Of Electrical And Electronics Engineers, 2018; OECD, 2019a; University of Montreal, 2017).[12] The UK Government's official public sector guide to safe and ethical AI has consolidated these into four "SUM values"—values that aim to *support*, *underwrite*, and *motivate* a responsible and reflective AI/ML innovation ecosystem and that are anchored in ethical concerns about human empowerment, interactive solidarity, individual and community wellbeing, and social justice (Leslie, 2019a). These are:

> **Respect** the dignity of individuals as persons:

> - Ensure the abilities of individuals to make free and well-informed decisions about their own lives
> - Safeguard their autonomy, their power to express themselves, and their right to be heard
> - Value the uniqueness of their aspirations, cultures, contexts, and forms of life
> - Secure their ability to lead a private life in which they are able to intentionally manage the transformative effects of the technologies that may influence and shape their development
> - Support their abilities to fully develop themselves and to pursue their passions and talents according to their own freely determined life plans

> **Connect** with each other sincerely, openly, and inclusively:
>
> - Safeguard the integrity of interpersonal dialogue and connection
> - Protect human interaction as a key for trust and empathy
> - Use technology to foster this capacity to connect so as to reinforce reciprocal responsibility and mutual understanding

> **Care** for the wellbeing of each and all:
>
> - Design and deploy AI to foster and to cultivate the welfare of all stakeholders whose interests are affected by their use

> - Do no harm with these technologies and minimize the risks of their misuse or abuse
> - Prioritize the safety and the mental and physical integrity of people when scanning horizons of technological possibility, conceiving of, and deploying AI applications

> **Protect** the priorities of justice, social values, and the public interest:
>
> - Treat all individuals equally and protect social equity
> - Use digital technologies to support the protection of fair and equal treatment under the law
> - Prioritize social welfare, public interest, and the consideration of the social and ethical impacts of innovation in determining the legitimacy and desirability of AI technologies
> - Use AI to empower and to advance the interests and well-being of as many individuals as possible
> - Think big-picture about the wider impacts of the AI technologies you are conceiving and developing. Think about the ramifications of their effects and externalities for others around the globe, for future generations, and for the biosphere as a whole

These SUM values form the basis of the Stakeholder Analysis component of the NHSx's *Code of Conduct for Data-Driven Health and Care Technology*. They are intended as a launching point for open and inclusive conversations about the individual and societal impacts of AI/ML innovation projects rather than to provide a comprehensive inventory of moral concerns and solutions. At the very outset of any AI/ML project, these should provide the normative point of departure for collaborative and anticipatory reflection, while, at the same time, allowing for the respectful and interculturally sensitive inclusion of other points of view.

It should be noted here that circumstances of a public health crisis such as the COVID-19 pandemic may place processes of deliberatively balancing and prioritizing conflicting or competing values under extreme pressure to yield decisions that generate difficult trade-offs between equally inviolable principles. In all cases, though there may be no a priori prescription or moral formula to determine such decisions in advance, research and innovation projects in data science must remain lawful and bound by obligations codified in existing international human rights agreements (WHO, 2016b) and data protection law (the General Data Protection Regulation, paradigmatically). In this respect, the *Siracusa Principles on the Limitation and Derogation Provisions in the International Covenant on Civil and Political Rights* (1984), provide a helpful reference point for considerations of the placement of permissible limitations on fundamental rights and freedoms in emergency situations where certain trade-offs are unavoidable to achieve legitimate objective interests. These affirm that any such restrictions should be a last resort after all other possible alternatives (which would have achieved the same outcome less intrusively) are exhausted, and that such restrictions should be legal, proportionate, reasonable, reviewable, evidence-based, and equitably executed (Boggio et al., 2008; Todrys et al., 2013).

Step IV: Generate and Cultivate Public Trust Through Transparency, Accountability, and Consent

The ultimate success of any AI/ML innovation project undertaken to combat COVID-19 will not only hang on the quality and performance of the product. It will also rest on whether or not a degree of public confidence in the safety and responsibility of the innovation has been established that is sufficient to foster its adoption by the health care community and society at large. Three key preconditions of trustworthy innovation deserve special attention.

First, all AI/ML innovation projects should proceed with *end-to-end transparency* to establish that design, discovery, and implementation *processes* have been undertaken responsibly and that *outcomes* are appropriately explainable and can be conveyed in plain language to all affected parties. Research undertaken to combat the SARS-CoV-2 outbreak should ceteris paribus occur as openly as possible. It should be carried out in a way that demonstrates to the public that innovation processes and products are ethically permissible as well as fair, safe, and worthy of trust (Association for Computing Machinery, 2017; Leslie 2019a). This entails the adoption of best practices mechanisms for responsible data sharing and for the assurance of data integrity such as the FAIR data and ALCOA plus principles mentioned above. Moreover, as our discussion of Wynants et al. (2020) suggested, research practices and methodological conduct should be carried out deliberately, transparently, and in accordance with recording protocols that enable the reproducibility and replicability of results. For prediction models, the documentation protocols presented in Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) is a good starting point for best conduct guidelines in reporting (Collins et al., 2015; Moons et al., 2015).

Researchers and innovators should likewise ensure that the results of their AI/ML models are reasonably and appropriately intelligible to users and affected individuals. Interpretable models and results will be a crucial factor in the adoption of AI/ML decision-support systems in clinical environments. They will also enable more effective and evidence-based assurance that AI/ML systems will operate safely, reliably, robustly, and equitably. Though this still remains something of a difficult issue for the complex, opaque classes of AI/ML algorithms, researchers and innovators should nevertheless prioritize the interpretability and explainability of their models from the start of their projects and, where applicable, maximize the accuracy and fidelity of any supplementary explanation methods they use to access the rationale of the complex models they deploy. They should also prioritize the use of interpretable methods, when structured data with meaningful representations are being utilized and pursue diligent techniques of iterative knowledge discovery as well as sufficient consultation with domain experts (Gilvary, 2019; Rudin, 2019).

A helpful reference point for this component of outcome transparency can be found in *Explaining Decisions Made with AI*, a guidance recently published by the United Kingdom's Information Commissioner's Office and the Alan Turing Institute. This guidance takes a holistic, end-to-end, and context-based approach to building AI/ML systems that are explainable-by-design. It focuses on the importance of tailoring both design-time and run-time strategies of producing understandable results to each model's specific use-case and practical context. This vital contextual aspect includes the specific subfield or area in which the clinical end-user operates, and the individual circumstances of the person receiving the decision. The guidance stresses a values-based approach to the governance of AI/ML explanations, presenting four principles of explainability that steer the recommendations it proposes: be transparent, be accountable, consider context, and reflect on impacts.

Building off these, it identifies a range of different explanation types, which cover various facets of an explanation, such as explanations of who is responsible, explanations of the rationale that led to a particular decision, explanations of how data has been collected, curated, and used, and explanations of measures taken across an AI/ML model's design and deployment to ensure fair and safe outcomes (Leslie & Cowls, 2020). Finally, it emphasizes that, in every individually impacting case, the statistical generalizations that underlie the rationale of any decision-support system's output should be translated into plain, socially meaningful language and applied by the end-user or implementer with due regard for the concrete life circumstance of the affected decision subject.

A second precondition of trustworthy innovation is accountability. All AI/ML innovation projects should proceed with *end-to-end accountability* to ensure both that humans are answerable for the parts they play across the entire AI/ML design, discovery, and implementation workflow and that the results of this work are traceable from start to finish. Diligent accountability protocols that are put in force across the AI/ML lifecycle will ensure public confidence that innovation processes prioritize patient and consumer interests from beginning to end. Members of civil society, domain experts, and other relevant stakeholders should also be included in the AI/ML workflow through the institution of independent advisory consortia, which function as sounding boards as well as sense-checks and oversight mechanisms throughout innovation processes.

Finally, these regimes of transparency and accountability should facilitate *informed community and individual consent* that reflects the contexts and reasonable expectations of affected stakeholders**.** Trust-building through community consultation should be utilized to foster the development of equal and respectful relationships—true partnerships—among researchers, health care professionals, and affected individuals and communities (Wright et al., 2020). Furthermore, public buy-in should come both from the groups in wider society that are impacted by the products of AI/ML innovation projects and from each individual who is directly affected by the use of these products. This can be achieved, on this broader scale, through effective ex ante public communication of the scope and nature of the AI/ML innovations undertaken. Such public engagement should provide nontechnical synopses of the research as well as summaries of the measures taken across the project lifecycle to ensure safe, ethical, equitable, and appropriately explainable outcomes.

Step V: Foster equitable innovation and protect the interests of the vulnerable.

Even before the COVID-19 pandemic, vulnerable and historically disadvantaged social groups were especially in peril of being harmed by or excluded from the benefits of data-driven technologies (Barocas & Selbst, 2016; Eubanks, 2018; Gianfranceso et al., 2018; Noble, 2018). Patterns of social inequity, marginalization, and injustice are often 'baked in' to the data distributions on which AI/ML systems learn. Over the past decade, a growing body of fairness-aware and bias-mitigating approaches to AI/ML design and use has been bringing many of these issues out into the open both in terms of academic research (for helpful surveys, see Friedler et al., 2019; Grgic-Hlaca et al., 2018; Mehrabi et al., 2019; Romei & Ruggieri, 2013; Verma & Rubin, 2018; Žliobaitė, 2017) and in terms of practically applicable user interfaces (several tools for fairness-aware design and bias auditing have been created, such as University of Chicago's Aequitas open source bias audit toolkit for machine learning developers, TU Berlin's Datasets and software for detecting algorithmic discrimination, and IBM's Fairness 360 open source toolkit).[13] This increasing focus on issues of bias and discrimination has brought needed attention to the deep-rooted dynamics of data set discrimination that are in peril of perpetuating many existing health inequities. Such dynamics have been evidenced, for example, in studies that have shown patterns of misclassified risk assessment of inherited cardiac conditions in Black Americans for reason of their lack of representation in genetic data sets (Manrai et al., 2016), information disparities across racial, ethnic, and ancestral subgroups about clinically relevant genetic variants in the Genome Aggregation Database (Popejoy et al., 2018), biased clinical risk assessment of atherosclerotic disease due to the overrepresentation of White patients in the Framingham Risk Factors cardiac evaluation tool (Gijsberts et al., 2015), and information gaps in the capacity of decision support tools to pick up diagnostic and treatment relevant signals in EHRs from vulnerable patient subgroups, who have irregular or limited access to health care (Arpey, 2017; Gianfancesco, 2018; Ng et al., 2017).

Though these instances highlight the importance of scrutinizing rapidly proliferating COVID-19 data sets for representativeness, balance, and inclusion of relevant information about all affected social groups across the demographic whole, the prevalence of health inequities they indicate call attention to other potential sources of pandemic-related digital discrimination. All-too-often, vulnerable or socioeconomically disadvantaged stakeholders are subject to material conditions, which make access to potentially beneficial digital technologies unavailable (Cahan et al., 2019; Weiss et al., 2018). Those who design digital apps used for contact tracing (and all other proposed mHealth tools and solutions) should pay special attention to those slices of the population where mobile smartphones are not used or unavailable for reasons of disadvantage, age, inequity, or other vulnerability. The burden is on policymakers, public health officials, data scientists, and AI/ML developers to come together with affected stakeholders to figure out how to include these potentially left-out members of our communities in consequential policies, initiatives, and innovations. If anything, this crisis should be an opportunity to critically assess and redress elements of the digital divide that still define so much of contemporary society and that help to perpetuate more widespread societal inequities.

In this respect, applied concepts of fairness and health equity should not simply be treated *in the abstract* as self-edifying ideals or ornaments of justice that can be engineered into AI/ML technologies through technical retooling or interpolation. This approach will produce a blindered range of vision whereby only the patterns of bias and discrimination in underlying data distributions that can be measured, formalized, and statistically digested are treated as worthy and actionable indicators of inequity, and this to exclusion of the subcutaneous sociocultural dynamics of domination that slip through cracks of quantification (Fazelpour & Lipton, 2020). Rather, the existing sociohistorical, economic, and political patterns and qualities of disadvantage that create material conditions of injustice must be taken as the starting point for reflection on the impacts and prospects of technological interventions. This means that the *terminus ad quem* of any and all attempts to protect the interests of the vulnerable through the mobilization of AI/ML innovation should be anchored in reflection on the concrete, bottom-up *circumstances of justice*, in its *historical and material preconditions*.

From this more pragmatic point of view (Dielman et al., 2017), there must be a prioritization of the real-world problems at the roots of lived injustice—problems that can then be treated as challenges "remediable" (Sen, 2011) by concerted social efforts and struggles for rectification, redistribution, and recognition (Fraser, 2010; Fraser & Honneth, 2003; Honneth, 2012). Only then will true-to-life demands for health equity and social justice be properly re-envisionable with and though the eyes of the oppressed. Only then will such demands become properly visible as struggles against the moral injuries inflicted by unjust social arrangements that obstruct the *participatory parity* of citizens in pursuing their unique paths to flourishing and in fully contributing to the moral and political life of the community.

5. Conclusion: Mobilizing Responsible AI Innovation to Help Today and to Shape the Society of Tomorrow

The ethical challenges faced by those innovators who are engaged in the second-front battle against COVID-19 have both immediate and intergenerational stakes. By carrying out their research and innovation ethically, transparently, and accountably, they will be better able to gain public trust, to accelerate collaborative problem-solving amid a global community of scientists, to support the evidence-based clinical judgments of overtaxed doctors, to ease the immense and growing socioeconomic hardships borne by most of present humanity, and to better prepare us for future pandemics.

But, these same innovators are confronted with dynamics of power and societal ills that together create conditions ripe for the abuse and misuse of the technological tools that they build and deploy. The data science and AI/ML community must therefore also act reflectively to safeguard cherished freedoms and values, the losses of which will very likely devastate our species for many generations to come. Taking this sort of anticipatory action is, however, well within its powers. Deliberate choices made, here and now, to engage in ethically informed and democratically governed innovation will not only help contemporary society build critical resistance to incipient strains of digital domination, it will facilitate the development of a future society that is more humane, rational, and enlightened.

For hundreds of years, at least since the 17th-century dawning of the Baconian and Newtonian revolutions in the natural sciences, the drive to improve the human lot through the fruits of scientific discovery has guided the steady, albeit imperfect, forward progress of socially responsible innovation. Led by the torchlight of social conscience and reason, this collaborative project has relied on inclusive, equitable, and democratic practices of research that have simultaneously served as a model for the participatory attainment of legitimate social arrangements and therefore for the freedom and openness of modern society itself. Being true to their practices, responsible scientific researchers must now draw on these progressive energies to help steward humankind through this troubled time and into a better, more empowering, and more just species life for the society of tomorrow.

Appendix

The Normative Dimension of Modern Scientific Advancement

Throughout this article, I have alluded to the normative dimension of the history of modern scientific advancement. Though a full elaboration of this is beyond the scope of the current endeavor, the topic is worthy of some brief clarification and expansion. By making explicit the moral grammar underlying the practical success of modern scientific methods, we can begin to better discern a path toward the realization of its beneficial potentials, while developing sightlines that will help us to steer clear of its greatest dangers. From this normative-historical perspective, the story of modern science is a story about how the successful development of a particular set of inclusive and consensus-based social practices of rational problem-solving carried out in the face of insuperable contingency has relied upon a corresponding release of the moral-practical potentials for cognitive humility, mutual responsibility, egalitarian reciprocity, individual autonomy, and unbounded social solidarity.[14]

As the broad-stroked narrative goes, at the very beginning of modernity, the deterioration of the religious and teleological order of things that typified traditional, premodern ways of life spurred the development of a thoroughgoing but salutary skepticism among a new generation of early modern scientists. The novel pressure to cope with the hardships of contingent reality without recourse to the authority of divine commandments or laws fixed by an intrinsically meaningful cosmic order (Taylor, 1989) consequently fueled an increasing awareness of the inescapable uncertainty that seemed to define the epistemic fragility, fallibility, and finitude of the human condition (Blumenberg, 1983). Such a starting point in a reflexive acknowledgement of self-limitation and 'learned ignorance' came to form the practical and epistemological basis of the experimental method of modern science (as initially exemplified in the work of pioneers like Pierre Gassendi, Francis Bacon, John Locke, and Isaac Newton).[15] This meant a shift from traditional modes of reasoning that appealed to the "inner nature of things and their necessary causes" to a new "science of experience" (Gassendi, 1624/1966) that was anchored in open-ended, collaborative problem-solving and carried out through ever-provisional forms of experimentation, reason-giving, and consensus-formation. In the midst of such a dramatic sociocultural sea-change, modern scientists became responsible to each other for sharing experience through standardized procedural mechanisms of rational warrant (like inductive reasoning and the experimental method) and for creating and reproducing the commonly held vocabularies that alone could shape the possibilities of their innovation and discovery.

Thinkers from Charles Sanders Peirce and John Dewey to Karl-Otto Apel, Robert Brandom, and Jürgen Habermas have long emphasized the importance of reconstructing the normative presuppositions that lie behind the emergence of these consensus-based and procedurally rational social practices. On this view, grasping the moral-practical enabling conditions for this new way of concerted human coping can help us to better comprehend the social and ethical determinants that have weighed heavily in the success of modern science itself. There are several.

The first is *the necessity of cognitive and methodological humility*. Dewey, in this connection, points out the "importance of uncertainty" (Dewey, 1933/1997, p. 12) and of the unending need to draw upon what Peirce (1877) called the "irritation of doubt" (p. 233) in the pursuit of open scientific inquiry, tentative suggestion, and experimentation. A starting point in cognitive and methodological humility functions as a lynchpin of the essential corrigibility and incompletability of modern scientific research and innovation. It secures the "unprejudiced openness that characterizes [its] cognitive process" (Habermas, 1992, p. 36). That no fallible interlocutor is entitled to have the last word in matters of scientific investigation leaves each participant in the unbounded community of inquiry (Apel, 1998; Peirce, 1868) no choice but to speak, to ask of others 'Why?,' to demand from them reasons for their claims and conclusions that are continuously liable to a mobile tribunal of ongoing rational assessment, criticism, and further observation (Leslie, 2016). The indeterminate and antiauthoritarian character of this open process of modern scientific inquiry unlocks trajectories of indefinite improvement at the same time as it lines up with the "inescapable incompleteness" (Rogers, 2009, p. xii) of modern democratic ways of life, which are legitimated by persistent practices of discursive exchange and meaning redemption. Directly drawing inspiration from "the spirit and method of science," Dewey summarizes, "the prime condition of a democratically organized public is a kind of knowledge and insight which does not yet exist… An obvious requirement is freedom of social inquiry and of distribution of its conclusions" (Dewey, 1927, p.166).

Already implicit in the practical concomitants of the demand for methodological humility is a second normative precondition for the success of modern science: *the imperative of publicity and the responsibilities of communication and listening*. An unbounded community of scientific inquiry can endure as such only insofar as it is organized around "free and systematic communication" (Dewey, 1927, p.167). The tentative and corrigible character of modern scientific insights makes this kind of publicity necessary inasmuch as inclusive debate and conversation are needed to ensure the continuous revision of beliefs and to foster the enlargement of an evolving space of scientific creativity and innovation (Mill, 1859/2006). This entails that "no one who could make a relevant contribution concerning a controversial validity claim must be excluded" (Habermas, 2008, p. 50). Likewise, all relevant positions, opinions, and information must be aired, exchanged, and weighed so that the stance participants take can be motivated "by the revisionary power of free-floating reasons" (Habermas, 2008, p. 50). The boundless conversation that underwrites the advancement of scientific inquiry must, along these lines, be open and accessible to all. Scientists have a responsibility to communicate their ideas plainly and to as wide an audience as possible, and nonscientist members of the public have a corollary responsibility to listen (Asimov, 1987; Leslie 2020).

A third normative precondition for the advancement of modern science stems from the *normative standing of participants involved in the ongoing rational dialogue of the scientific "communication community"* (Apel, 1998, p. 225). In order for those engaged in practices of giving and asking for reasons *to be conferred the authority* to endorse validity claims and, in turn, *to be held accountable* for their commitments to these, they must reciprocally grant each other normative status as being rational and responsible agents (Brandom, 1994/2001, 2000, 2013). The mutual conferral of this normative standing operates as a pragmatic presupposition of communicative practices of scientific inquiry, for it makes interlocutors liable to each other for the rational assessment of the arguments they tender. Beyond this, the process of rational assessment itself entails a further set of procedural requirements that place additional normative-pragmatic demands on participants engaged in inquiry. Because the claims to propositional truth that are building blocks of modern science carry an unconditional, context-bursting force that reaches beyond the embodied and factually situated circumstances in which they are uttered, these claims structurally mandate procedures that, at once, "guarantee the impartiality of the process of judging" (Habermas, 1990/2001, p. 122) and secure an "egalitarian universalism" (Habermas, 2008, p. 49) in the practices of giving and asking for reasons by which propositions gain rational acceptability. Chief among such unavoidable idealizing suppositions of those engaged in rational discourse are mutual respect, egalitarian reciprocity, equal right to engage in communication and equal opportunity to contribute to it, noncoercion, participatory parity, and sincerity (Habermas, 1990/2001, 1992, 1996/2002, 1998/2003, 2008).

A final normative precondition, *the intrinsic sociality of science*, is predicated on the role that scientific inquiry plays as a practical medium of problem-solving through collaboration, reason-giving, and experimentation. Although the explanatory ambitions of the modern natural sciences have largely been anchored in making truth claims about the world through physical observations, quantified measurements, and the experimental practices of hypothesis testing, these approaches are, at bottom, rooted in social processes of intersubjective communication that are driven by shared endeavors to cope with challenges deemed worthy of response. Scientific practices are always already embedded in a community of interpretation and in holistic contexts of individual life plans and collective social projects (Apel, 1998; Royce, 1908/1995). The evolution of scientific inquiry occurs within a changing space of reasons, interpretations, and values (Apel, 1984, 1999; Sellars, 1956/1997; Taylor, 1964/1980; Von Wright, 1971/2004). And, as humans adjust their purposes and goals to meet the needs of their times, science too changes its focus, outlook, and direction. This holistic and value-oriented departure point of scientific practices implies that modern science should not be viewed, first and foremost, as operationally independent from human beliefs, aims, and interpretations, but rather as an ethically implicated set of problem-solving practices that are steered by the values and commitments of its embodied producers. In Dewey's words, "The notion of the complete separation of science from the social environment is a fallacy which encourages irresponsibility, on the part of scientists, regarding the social consequences of their work" (Dewey, 1938, p. 489). This intrinsic sociality of science functions then as an enabling condition of the responsibility of innovation and of its humane pursuit of what Francis Bacon called the "relief man's estate [through discovery]" (Bacon, 1605/2001).

To close here, it may be useful to note that, taken together, these normative-pragmatic presuppositions of the advancement of modern science continuously push researchers and innovators to think beyond themselves and their existing communities of practice to consider their role in safeguarding the endurance of a greater living whole. Prompted to see themselves in this light, they are better equipped to embrace the essential positions they occupy both as stakeholders vested in a world-yet-to-come and as committed members of two broader, expanding circles. Their participation in the first of these involves playing an active, albeit transient, part in an unbounded community of learning and discovery that is charged with advancing the "permanent interests of humankind as a progressive being," to paraphrase J.S. Mill (1859/2006). The execution of such a species-level commission to improve the present and future conditions of life demands that, not only scientists, but all members of humanity be able to carry out the indefinite and transgenerational tasks of shared knowledge-creation and collaborative world-making through unfettered communication, consensus-based value articulation, and deliberative will formation. To this end, humankind itself must become ever more capable of inclusively cultivating and drawing upon the unique talents, passions, and callings of each of its increasing number. That is, as this circle of shared learning and discovery expands, every human being should be capacitated to pursue their own path to intellectual and creative self-realisation so that the universal fulfilment of the full potential of each can usher forward the greater social project of the sustenance and flourishing of all. This civilizational impetus to "fully integrated personality" (Dewey, 1946, p. 148) entails that any arbitrary socioeconomic or geopolitical barriers to equitable flourishing be demolished so that no future Ramanujan, Curie, Turing, or Einstein can be lost to the dynamics of societal oppression that stamp out the flames of human genius before they have a chance to ignite.

The second expanding circle in which researchers and technologist are included extends the responsibility of innovation to all members of the circle of life itself. Outfitted with multiplying technological capacities to bring about species self-annihilation, mass extinction, and biospheric catastrophe, the human community of learning and discovery now finds itself implicated as a potentially cataclysmic force of nature, in its own right. Those at the tiller of scientific research and innovation are consequently no longer entitled to simply assume a kind of legitimate epistemological or ontological division between 'nature' and 'society.' The presumption of such a 'great divide' between natural and cultural worlds has promoted a misdirected self-perception among scientists that they are engaged in a neutral and value-free enterprise, thereby enabling reckless strains of the modern natural sciences to claim functional immunity from the curbing modes of ethical critique that derive from social environments in which they are situated. It has also allowed them to instrumentally treat the living and inanimate constituents of the natural world simply as objects available for appropriation, calculation, and control. With the ushering in of the Anthropocene epoch such a presumed dichotomy between nature and society has become increasingly implausible inasmuch as the scope of the anthropogenic impacts on climate and biosphere increasingly inculpates humankind not only as a natural force of geohistorical consequence but as an essential co-originator of the conditions of possibility for the survival and flourishing of life on earth. In this way, nature as such can no longer be seen merely as an object to be measured and manipulated but is now as us, above all, implicated as a subject bound by ethical obligations of existential import and intergenerational reach.

It follows from all this that the human community is part and parcel of a wider circle of natural organisms whence, over a fortunate 3.7-billion-year trajectory of evolutionary transformation, it has developed the exceptional capabilities for limitless creation and mass destruction for which it is, in the end, uniquely accountable. We should take cognizance here, however, that, although modern science has been an essential catalyst in facilitating the enabling conditions of these dangerous competences, it has concurrently shown a path to the societal acceptance of such a unique responsibly by spurring an ethical self-understanding of humanity's place in exactly this deep history of life. From Hutton and Lyell to Sedgwick and Darwin, modern scientific insights have enabled chastening and worldview de-centering access to a widening temporal frame of geological and evolutionary history within which the human species has been placed on a living continuum extending from the very first unicellular organisms to the shrinking plurality of flora and fauna that typifies our current era of biodiversity drain and the humanly prompted "sixth extinction". From the sharp end of the deep historical arc at which we now find ourselves, it is possible to peer back across thousands of millennia so to see that all biological individuals have been interlinked in such an evolving circle of life from the outset and that, notwithstanding the contemporary anthropogenic mass extermination of species, living matter's diversifying impetus has tracked the development of a kind a holistic unity within the unbounded community of the biospheric whole.

From the vista of humankind's membership in this broader biotic totality, our story may well be seen as a tale of two species. For, on the one hand, we are a species principally unconstrained in its pursuit of the boundless possibilities opened by the infinite generativity of its capacity for language, representation, and symbolic experience—a species which is, for precisely that reason, readily capacitated to affect the self-annihilation of planetary life—a dangerous species. On the other, we are a species endowed with recourse to the media of collaboration, communication, and criticism by means of which we are able to constrain our penchant for techno-scientific hubris and to sustain the futurity and flourishing of the greater living whole. We are an ethical species endowed with a deep historically ingrained sense of responsibility to the intrinsic worth of life as such and hence capable of stewarding the sustenance of the biosphere as its trustees and as its guardians. It is perhaps the greatest redeeming power of modern scientific advancement that it has granted us the wherewithal to tell this second, moral story.

Disclosure Statement

The author has no conflicts of interest to declare.

Acknowledgements

Not so many moons ago, this paper began as a short blog, written for the Alan Turing Institute, as the UK entered into COVID-19 lockdown. That initial piece quickly grew tight in its own digital skin. As it developed into a longer and more academic article, which aimed to reach a general audience but remain pertinent for specialists, thoughtful and incisive input from four anonymous reviewers and the editor-in-chief at the HDSR stewarded many improvements, though the deficiencies that remain in it are its author's responsibility alone. I would like to thank them as well as my research assistant, Christina Hitrova, for her efforts in helping me get together the bibliography. Corianna Moffatt's assistance with copyediting and formatting also proved invaluable as did helpful comments from Christopher Burr and Suzanne Smith, who both worked through the paper as it neared its completion.

References

Academy of Medical Royal Colleges. (2019). *Artificial intelligence in healthcare*. Academy of Medical Royal Colleges. https://www.aomrc.org.uk/wp-content/uploads/2019/01/Artificial_intelligence_in_healthcare_0119.pdf

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 559–560). https://doi.org/10.1145/3233547.3233667

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., & Xia, L. (2020). Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 Cases. *Radiology*, *295*(3), Article 200642. https://doi.org/10.1148/radiol.2020200642

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Al-qaness, M. A. A., Ewees, A. A., Fan, H., & Abd El Aziz, M. (2020). Optimization method for forecasting confirmed cases of COVID-19 in China. *Journal of Clinical Medicine*, *9*(3), 674. https://doi.org/10.3390/jcm9030674

Altman, D. G., Vergouwe, Y., Royston, P., & Moons, K. G. M. (2009). Prognosis and prognostic research: Validating a prognostic model. *BMJ*, *338,* Article b605. https://doi.org/10.1136/bmj.b605

Altuwaiyan, T., Hadian, M., & Liang, X. (2018). EPIC: Efficient Privacy-Preserving Contact Tracing for Infection Detection. *2018 IEEE International Conference on Communications (ICC)*, 1–6. https://doi.org/10.1109/ICC.2018.8422886

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *ArXiv:1806.07538 [Cs, Stat]*. http://arxiv.org/abs/1806.07538

Amoore, L. (2009). Algorithmic war: Everyday geographies of the war on terror. *Antipode*, *41*(1), 49–69. https://doi.org/10.1111/j.1467-8330.2008.00655.x

Amoore, L., & Raley, R. (2017). Securing with algorithms: Knowledge, decision, sovereignty. *Security Dialogue*, *48*(1), 3–10. https://doi.org/10.1177/0967010616680753

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research, 18,* 1–78.

Apel, K.-O. (1984). *Understanding and explanation: A transcendental-pragmatic perspective*. MIT Press.

Apel, K.-O. (1998). *Towards a transformation of philosophy*. Marquette University Press.

Apel, K.-O. (1999). *From a transcendental-semiotic point of view*. Manchester University Press; St. Martin's Press.

Apple. (2020, April 10). *Apple and Google partner on COVID-19 contact tracing technology*. Apple Newsroom. https://www.apple.com/uk/newsroom/2020/04/apple-and-google-partner-on-COVID-19-contact-tracing-technology/

Arpey, N. C., Gaglioti, A. H., & Rosenbaum, M. E. (2017). How socioeconomic status affects patient perceptions of health care: A qualitative study. *Journal of Primary Care & Community Health, 8*(3), 169–175. https://doi.org/10.1177/2150131917697439

Ashmore, R., Calinescu, R., & Paterson, C. (2019). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *arXiv preprint arXiv:1905.04223*.

Asimov, I. (1987). *Asimov's new guide to science* (Rev. ed). Penguin Books.

Association for Computing Machinery U.S. Public Policy Council. (2017). *Statement on algorithmic transparency and accountability*. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Bacon, F. (2001). *The advancement of learning* (pbk. ed). Modern Library. (Original work published 1605)

Bailly, S., Meyfroidt, G., & Timsit, J.-F. (2018). What's new in ICU in 2050: Big data and machine learning. *Intensive Care Medicine*, *44*(9), 1524–1527. https://doi.org/10.1007/s00134-017-5034-3

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around procedural privacy protections. *Communications of the ACM*, *57*(11), 31–33. https://doi.org/10.1145/2668897

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *ArXiv:1704.05796 [Cs]*. http://arxiv.org/abs/1704.05796

Bauer, P. C., & Freitag, M. (2018). Measuring trust. In E. M. Uslaner (Ed.), *The Oxford Handbook of Social and Political Trust* (pp. 15–36). Oxford University Press.

Bay, J., Kek, J., Tan, A., Hau, C. S., Yongquan, L., Tan, J., & Quy, T. A. (2020). *BlueTrace: A privacy-preserving protocol for community-driven contact tracing across borders*. https://bluetrace.io/static/bluetrace_whitepaper-938063656596c104632def383eb33b3c.pdf

Beauchamp, T. L., & Childress, J. F. (2013). *Principles of biomedical ethics* (7th ed). Oxford University Press.

Beck, B. R., Shin, B., Choi, Y., Park, S., & Kang, K. (2020). *Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV), Wuhan, China through a drug-target interaction deep learning model* [Preprint]. Microbiology. https://doi.org/10.1101/2020.01.31.929547

Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., and the CAMELYON16 Consortium, Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., … Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), Article 2199. https://doi.org/10.1001/jama.2017.14585

Bell, J., Butler, D., Hicks, C., & Crowcroft, J. (2020). TraceSecure: Towards privacy preserving contact Tracing. *ArXiv:2004.04059 [Cs]*. http://arxiv.org/abs/2004.04059

Berikol, G. B., Yildiz, O., & Özcan, İ. T. (2016). Diagnosis of acute coronary syndrome with a support vector machine. *Journal of Medical Systems*, *40*(4), Article 84. https://doi.org/10.1007/s10916-016-0432-6

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. *ArXiv:1703.09207 [Stat]*. http://arxiv.org/abs/1703.09207

Berke, A., Bakker, M., Vepakomma, P., Larson, K., & Pentland, A. "Sandy." (2020). Assessing disease exposure risk with location data: A proposal for cryptographic preservation of privacy. *ArXiv:2003.14412 [Cs]*. http://arxiv.org/abs/2003.14412

Berman, F., & Crosas, M. (2020). The research data alliance: Benefits and challenges of building a community organization. *Harvard Data Science Review, 2*(1). https://doi.org/10.1162/99608f92.5e126552

Bezuidenhout, L. M., Leonelli, S., Kelly, A. H., & Rappert, B. (2017). Beyond the digital divide: Towards a situated approach to open data. *Science and Public Policy*, *44*(4), 464–475. https://doi.org/10.1093/scipol/scw036

Bichindaritz, I., & Marling, C. (2006). Case-based reasoning in the health sciences: What's next? *Artificial Intelligence in Medicine*, *36*(2), 127–135. https://doi.org/10.1016/j.artmed.2005.10.008

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, *5*(4), 2403–2424. https://doi.org/10.1214/11-AOAS495

Bloomfield, R., & Netkachova, K. (2014). Building blocks for assurance cases. In *2014 IEEE International Symposium on Software Reliability Engineering Workshops*(pp. 186-191). IEEE.

Bloomfield, R., & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future–an Adelard perspective. In *Making Systems Safer* (pp. 51-67). Springer, London.

Blumenberg, H. (1983). *The legitimacy of the modern age* (1st. pbk. ed.). MIT Press.

BMA. (2020). *COVID-19–ethical issues. A guidance note*. https://www.bma.org.uk/media/2226/bma-COVID-19-ethics-guidance.pdf

Boberg, S., Quandt, T., Schatto-Eckrodt, T., & Frischlich, L. (2020). Pandemic populism: Facebook pages of alternative news media and the corona crisis—A computational content analysis.

ArXiv:2004.02566 [Cs]. http://arxiv.org/abs/2004.02566

Boggio, A., Zignol, M., Jaramillo, E., Nunn, P., Pinet, G., & Raviglione, M. (2008). Limitations on human rights: Are they justifiable to reduce the burden of TB in the era of MDR- and XDR-TB? *Health and Human Rights*, *10*(2), 121–126. https://doi.org/10.2307/20460107

Bolin, B., & Kurtz, L. C. (2018). Race, class, ethnicity, and disaster vulnerability. In H. Rodríguez, W. Donner, & J. E. Trainor (Eds.), *Handbook of disaster research* (pp. 181–203). Springer.

Bologna, G., & Hayashi, Y. (2017). Characterization of symbolic rules embedded in deep DIMLP networks: A challenge to transparency of deep learning. *Journal of Artificial Intelligence and Soft Computing Research*, *7*(4), 265–286. https://doi.org/10.1515/jaiscr-2017-0019

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Bouwmeester, W., Zuithoff, N. P. A., Mallett, S., Geerlings, M. I., Vergouwe, Y., Steyerberg, E. W., Altman, D. G., & Moons, K. G. M. (2012). Reporting and methods in clinical prediction research: A systematic review. *PLoS Medicine*, *9*(5), Article e1001221. https://doi.org/10.1371/journal.pmed.1001221

Brandom, R. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.

Brandom, R. (2001). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press. (Original work published 1994)

Brandom, R. (2013). *Reason in philosophy: Animating ideas*. Harvard University Press.

Brack, S., Reichert, L., & Scheuermann, B. (2020). *Decentralized contact tracing using a DHT and blind signatures*. International Association for Cryptologic Research. https://eprint.iacr.org/2020/398.pdf

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–215.

Briand, S., Mounts, A., & Chamberland, M. (2011). Challenges of global surveillance during an influenza pandemic. *Public Health*, *125*(5), 247–256. https://doi.org/10.1016/j.puhe.2010.12.007

Brisimi, T. S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I. Ch., & Shi, W. (2018). Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics*, *112*, 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007

Bull, S., Cheah, P. Y., Denny, S., Jao, I., Marsh, V., Merson, L., Shah More, N., Nhan, L. N. T., Osrin, D., Tangseefa, D., Wassenaar, D., & Parker, M. (2015). Best practices for ethical sharing of individual-level health research data from low- and middle-income settings. *Journal of Empirical Research on Human Research Ethics*, *10*(3), 302–313. https://doi.org/10.1177/1556264615594606

Bullock, J., Luccioni, A., Pham, K. H., Lam, C. S. N., & Luengo-Oroz, M. (2020). Mapping the landscape of artificial intelligence applications against COVID-19. *ArXiv:2003.11336 [Cs]*. http://arxiv.org/abs/2003.11336

Burgelman, J.-C., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., & Schouppe, M. (2019). Open science, open data, and open scholarship: European policies to make science fit for the Twenty-First Century. *Frontiers in Big Data*, *2*, Article 43. https://doi.org/10.3389/fdata.2019.00043

Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., & Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, *8*(1), Article 3395. https://doi.org/10.1038/s41598-018-21758-3

Cahan, E. M., Hernandez-Boussard, T., Thadaney-Israni, S., & Rubin, D. L. (2019). Putting the data before the algorithm in big data addressing personalized healthcare. *Npj Digital Medicine*, *2*(1), Article 78. https://doi.org/10.1038/s41746-019-0157-2

Calders, T., & Žliobaite, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Studies in applied philosophy, epistemology and rational ethics* (pp. 43–57). Springer. https://www.researchgate.net/profile/Bart_Custers3/publication/278661450_What_Is_Data_Mining_and_How_Does_It_Work/links/5b924523299bf147391feb30/What-Is-Data-Mining-and-How-Does-It-Work.pdf#page=58

Calvo, R. A., Deterding, S., & Ryan, R. M. (2020). Health surveillance during COVID-19 pandemic. *BMJ*, Article m1373. https://doi.org/10.1136/bmj.m1373

Canetti, R., Trachtenberg, A., & Varia, M. (2020). Anonymous collocation discovery: Harnessing privacy to tame the coronavirus. *ArXiv:2003.13670 [Cs]*. http://arxiv.org/abs/2003.13670

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Chan, J., Foster, D., Gollakota, S., Horvitz, E., Jaeger, J., Kakade, S., Kohno, T., Langford, J., Larson, J., Singanamalla, S., Sunshine, J., & Tessaro, S. (2020). PACT: Privacy sensitive protocols and mechanisms for mobile contact tracing. *ArXiv:2004.03544 [Cs]*. http://arxiv.org/abs/2004.03544

Chatila, R., & Havens, J. C. (2019). The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (Vol. 95, pp. 11–16). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_2

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2019). This looks like that: Deep learning for interpretable image recognition. *ArXiv:1806.10574 [Cs, Stat]*. http://arxiv.org/abs/1806.10574

Chen, E., Lerman, K., & Ferrara, E. (2020). COVID-19: The first public coronavirus Twitter dataset. *ArXiv:2003.07372 [Cs, q-Bio]*. http://arxiv.org/abs/2003.07372

Cho, H., Ippolito, D., & Yu, Y. W. (2020). Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *ArXiv:2003.11511 [Cs]*. http://arxiv.org/abs/2003.11511

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2017). RETAIN: An Interpretable predictive model for healthcare using reverse time attention mechanism. *ArXiv:1608.05745 [Cs]*. http://arxiv.org/abs/1608.05745

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *ArXiv:2003.05004 [Nlin, Physics:Physics]*. http://arxiv.org/abs/2003.05004

CoEpi: Community Epidemiology in Action. (2020). https://www.coepi.org/

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., & Altman, D. G. (2014). External validation of multivariable prediction models: A systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1), Article 40. https://doi.org/10.1186/1471-2288-14-40

Collins, G. S., Omar, O., Shanyinde, M., & Yu, L.-M. (2013). A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *Journal of Clinical Epidemiology*, *66*(3), 268–277. https://doi.org/10.1016/j.jclinepi.2012.06.020

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD). *Circulation, 131* (2), 211–219. https://doi.org/10.1161/CIRCULATIONAHA.114.014508

Community, T. T. W., Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., & Whitaker, K. (2019). *The Turing way: A handbook for reproducible data science*. Zenodo. https://doi.org/10.5281/ZENODO.3233986

COVIDWatch. (2020). https://www.COVID-watch.org/

Crane, J. (2011). Scrambling for Africa? Universities and global health. *The Lancet*, *377*(9775), 1388–1390. https://doi.org/10.1016/S0140-6736(10)61920-4

Creemers, R. (2018). China's social credit system: An evolving practice of control. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3175792

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, *5*(2), 120–134. https://doi.org/10.1089/big.2016.0048

Damen, J. A. A. G., Hooft, L., Schuit, E., Debray, T. P. A., Collins, G. S., Tzoulaki, I., Lassale, C. M., Siontis, G. C. M., Chiocchia, V., Roberts, C., Schlüssel, M. M., Gerry, S., Black, J. A., Heus, P., van der Schouw, Y. T., Peelen, L. M., & Moons, K. G. M. (2016). Prediction models for cardiovascular disease risk in the general population: Systematic review. *BMJ*, Article i2416. https://doi.org/10.1136/bmj.i2416

Danquah, L. O., Hasham, N., MacFarlane, M., Conteh, F. E., Momoh, F., Tedesco, A. A., Jambai, A., Ross, D. A., & Weiss, H. A. (2019). Use of a mobile application for Ebola contact tracing and monitoring in northern Sierra Leone: A proof-of-concept study. *BMC Infectious Diseases*, *19*(1), Article 810. https://doi.org/10.1186/s12879-019-4354-z

Dash, S., Günlük, O., & Wei, D. (2018). Boolean decision rules via column generation. *ArXiv:1805.09901 [Cs]*. http://arxiv.org/abs/1805.09901

Davidson, H. (2020, April 1). China's coronavirus health code apps raise concerns over privacy. *The Guardian*. theguardian.com/world/2020/apr/01/chinas-coronavirus-health-code-apps-raise-concerns-over-privacy

de Montjoye, Y.-A. (2015). *Computational privacy: Towards privacy-conscientious uses of metadata* [Master's thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/101850

Denney, E., Pai, G., & Habli, I. (2015). Dynamic safety cases for through-life safety assurance. *IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 2, pp. 587-590). IEEE.

Department of Health & Social Care. (2019). *Code of conduct for data-driven health and care technology* [Guidance]. https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology

Department of Health, Education, and Welfare, & National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1974). The Belmont report. Ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists*, *81*(3), 4–13.

Desai, T., Ritchie, F., & Welpton, R. (2016). *Five safes: Designing data access for research.* (No. 1601; Economics Working Paper Series). University of the West of England. https://www2.uwe.ac.uk/faculties/bbs/Documents/1601.pdf

Dewey, J. (1927). *The public and its problems: An essay in political inquiry*. Swallow Press.

Dewey, J. (1938). *Logic the theory of inquiry*. Henry Holt and Company.

Dewey, J. (1997). *How we think*. Dover Publications. (Original work published 1933)

Dieleman, S., Rondel, D., & Voparil, C. J. (Eds.). (2017). *Pragmatism and justice*. Oxford University Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv:1702.08608 [Cs, Stat]*. http://arxiv.org/abs/1702.08608

Dove, E. S. (2015). Reflections on the concept of open data. *SCRIPTed*, *12*(2), 154–166. https://doi.org/10.2966/script.120215.154

Dupré, J. (1993). *The disorder of things: Metaphysical foundations of the disunity of science*. Harvard University Press.

Eden, G., Jirotka, M., & Stahl, B. (2013). Responsible research and innovation: Critical reflection into the potential social consequences of ICT. *IEEE 7th International Conference on Research Challenges in Information Science (RCIS)*, 1–12. https://doi.org/10.1109/RCIS.2013.6577706

Ehsani-Moghaddam, B., Martin, K., & Queenan, J. A. (2019). Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. *Health Information Management Journal*, Article 183335831988774. Advance online publication. https://doi.org/10.1177/1833358319887743

Engineering & Physical Sciences Research Council. (2011). *Principles of robotics: Regulating robots in the real world.* https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Engineering & Physical Sciences Research Council. (2013). *Framework for Responsible Innovation*. https://epsrc.ukri.org/research/framework/

Erikson, E. H. (1959). *Identity and the life cycle: Selected papers*. International Universities Press.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

EU High-Level Expert Group on Artificial Intelligence. (2019, April 3). *Ethics Guidelines for Trustworthy AI* [Text]. FUTURIUM–European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

Eubanks, V. (2018). *Automating inequality*. St. Martin's Press.

European Commission. (2014). *Consultation on "Science 2.0": Science in Transition*. Research & Innovation (European Commission). https://ec.europa.eu/research/consultations/science2.0/consultation_en.htm

Fecher, B., & Friesike, S. (2013). Open science: One term, five schools of thought. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2272036

Fergus, P., Hussain, A., Al-Jumeily, D., Huang, D.-S., & Bouguila, N. (2017). Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms. *BioMedical Engineering OnLine*, *16*(1), Article 89. https://doi.org/10.1186/s12938-017-0378-z

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., & Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, *368*(6491), Article eabb6936. https://doi.org/10.1126/science.abb6936

Ferryman, K., & Pitcan, M. (2018). *Fairness in precision medicine* (p. 54). Data & Society. https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf

Fitzsimons, J. K., Mantri, A., Pisarczyk, R., Rainforth, T., & Zhao, Z. (2020). A note on blind contact tracing at scale with applications to the COVID-19 pandemic. *ArXiv:2004.05116 [Cs]*. http://arxiv.org/abs/2004.05116

Floridi, L. (2010). *The Cambridge handbook of information and computer ethics*. Cambridge University Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=501402

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, *32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., & Lord Clement-Jones. (2019, March 20). The five principles key to any ethical framework for AI. *New Statesman*. https://tech.newstatesman.com/policy/ai-ethics-framework

Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research—Commentary. *BioMedical Engineering OnLine*, *13*(1), Article 94. https://doi.org/10.1186/1475-925X-13-94

Fothergill, A., & Peek, L. A. (2004). Poverty and disasters in the United States: A review of recent sociological findings. *Natural Hazards*, *32*(1), 89–110. https://doi.org/10.1023/B:NHAZ.0000026792.76181.d9

Foucault, M. (1988). *Madness and civilization: A history of insanity in the age of reason* (Vintage Books Ed.). Random House. (Original work published 1961).

Foucault, M. (2007). *The order of things: An archaeology of the human sciences*. Routledge. (Original work published 1966).

Fourcade, M., & Healy, K. (2013). Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society*, *38*(8), 559–572. https://doi.org/10.1016/j.aos.2013.11.002

Fourcade, M., & Healy, K. (2016). Seeing like a market. *Socio-Economic Review*, *15*(1), Article mww033. https://doi.org/10.1093/ser/mww033

Fraser, N. (2010). *Scales of justice: Reimagining political space in a globalizing world*. Polity Press.

Fraser, N., & Honneth, A. (2003). *Redistribution or recognition? A political-philosophical exchange*. Verso.

Freitas, A. A. (2013). *Comprehensible classification models–A position paper*. *15*(1), 10.

French, M., & Monahan, T. (2020). Dis-ease surveillance: How might surveillance studies address COVID-19. *Surveillance & Society*, *18*(1). https://ojs.library.queensu.ca/index.php/surveillance-and-society/index

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 329–338. https://doi.org/10.1145/3287560.3287589

Fuchs, C. (2010). Labor in informational capitalism and on the internet. *The Information Society*, *26*(3), 179–196. https://doi.org/10.1080/01972241003712215

Fussell, S., & Knight, W. (2020, April 14). The Apple-Google contact tracing plan won't stop COVID alone. *Wired*. https://www.wired.com/story/apple-google-contact-tracing-wont-stop-COVID-alone/

Galison, P., & Stump, D. J. (Eds.). (1996). *The sisunity of science: Boundaries, contexts, and power*. Stanford University Press.

Gassendi, P. (1966). Exercitationes paradoxicae adversus Aristoteleos. In *Philosophy of the Sixteenth and Seventeenth Centuries*. Free Press. (Original work published 1624)

Ge, X., Rijo, R., Paige, R. F., Kelly, T. P., & McDermid, J. A. (2012). Introducing goal structuring notation to explain decisions in clinical practice. *Procedia Technology*, *5*, 686-695.

Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., Yang, H., Hong, L., Wu, N., Yuan, E., Cheng, L., Lei, Y., Shu, H., Feng, X., Jiang, Z., Chi, Y., Guo, X., Cui, L., Xiao, L., ... Zeng, J. (2020). *A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19* [Preprint]. Systems Biology. https://doi.org/10.1101/2020.03.11.986836

Ghassemi, M., Naumann, T., Schulam, P., Chen, I. Y., & Ranganath, R. (2018). *A review of challenges and opportunities in machine learning for health*. https://arxiv.org/pdf/1806.00388.pdf

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, *178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

Gilvary, C., Madhukar, N., Elkhader, J., & Elemento, O. (2019). The missing pieces of artificial intelligence in medicine. *Trends in Pharmacological Sciences*, *40*(8), 555–564. https://doi.org/10.1016/j.tips.2019.06.001

Gijsberts, C. M., Groenewegen, K. A., Hoefer, I. E., Eijkemans, M. J. C., Asselbergs, F. W., Anderson, T. J., Britton, A. R., Dekker, J. M., Engström, G., Evans, G. W., de Graaf, J., Grobbee, D. E., Hedblad, B., Holewijn, S., Ikeda, A., Kitagawa, K., Kitamura, A., de Kleijn, D. P. V., Lonn, E. M., ... den Ruijter, H. M. (2015). Race/ethnic differences in the associations of the Framingham Risk Factors with carotid IMT and cardiovascular events. *PLOS ONE*, *10*(7), Article e0132321. https://doi.org/10.1371/journal.pone.0132321

Goldacre, B., Harrison, S., Mahtani, K. R., & Heneghan, C. (2015). *WHO consultation on data and results sharing during public health emergencies* [Background briefing]. Centre for Evidence-Based Medicine, Nuffield Department of Primary Care Health Sciences, University of Oxford. https://www.who.int/medicines/ebola-reatment/background_briefing_on_data_results_sharing_during_phes.pdf

Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society - WPES '06*, 77. https://doi.org/10.1145/1179601.1179615

Google. (2020, April 10). *Apple and Google partner on COVID-19 contact tracing technology*. https://blog.google/inside-google/company-announcements/apple-and-google-partner-COVID-19-contact-tracing-technology/

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., Bernheim, A., & Siegel, E. (2020). Rapid AI development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning CT image analysis. *ArXiv:2003.05037 [Cs, Eess]*. http://arxiv.org/abs/2003.05037

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2018). *Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning*. Association for the Advancement of Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16523

Grimpe, B., Hartswood, M., & Jirotka, M. (2014). Towards a closer dialogue between policy and practice: Responsible design in HCI. *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, 2965–2974. https://doi.org/10.1145/2556288.2557364

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Gunning, D. (2017). Explainable artificial intelligence (XAI). *DARPA*, 36. Defense Advanced Research Projects Agency. https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Ha, Y. P., Tesfalul, M. A., Littman-Quinn, R., Antwi, C., Green, R. S., Mapila, T. O., Bellamy, S. L., Ncube, R. T., Mugisha, K., Ho-Foster, A. R., Luberti, A. A., Holmes, J. H., Steenhoff, A. P., & Kovarik, C. L. (2016). Evaluation of a mobile health approach to tuberculosis contact tracing in Botswana. *Journal of Health Communication*, *21*(10), 1115–1121. https://doi.org/10.1080/10810730.2016.1222035

Habermas, J. (1992). *Postmetaphysical thinking: Philosophical essays*. MIT Press.

Habermas, J. (2001). *Moral consciousness and communicative action*. MIT Press. (Original work published 1990)

Habermas, J. (2002). *The inclusion of the other: Studies in political theory*. Polity. (Original work published 1996)

Habermas, J. (2003). *On the pragmatics of communication*. Polity. (Original work published 1998)

Habermas, J. (2008). *Between naturalism and religion: Philosophical essays*. Polity Press.

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization*, *98*(4), 251–256. https://doi.org/10.2471/BLT.19.237487

Hays, R., & Daker-White, G. (2015). The care.data consensus? A qualitative analysis of opinions expressed on Twitter. *BMC Public Health*, *15*(1), Article 838. https://doi.org/10.1186/s12889-015-2180-9

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Health Foundation (2012). Evidence: Using safety cases in industry and healthcare. https://www.health.org.uk/sites/default/files/UsingSafetyCasesInIndustryAndHealthcare.pdf

Heinrich, B., Kaiser, M., & Klier, M. (2007, July). *Metrics for measuring data quality—Foundations for an economic data quality management*. 2nd International Conference on Software and Data Technologies (ICSOFT). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.699.7650&rep=rep1&type=pdf

Hekmati, A., Ramachandran, G., & Krishnamachari, B. (2020). CONTAIN: Privacy-oriented contact tracing protocols for epidemics. *ArXiv:2004.05251 [Cs]*. http://arxiv.org/abs/2004.05251

Hellström, T. (2003). Systemic innovation and risk: Technology assessment and the challenge of responsible innovation. *Technology in Society*, *25*(3), 369–384. https://doi.org/10.1016/S0160-791X(03)00041-1

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., Lehmann, H. P., Hripcsak, G., Hartzog, T. H., Cimino, J. J., & Saltz, J. H. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research: *Medical Care*, *51*, S30–S37. https://doi.org/10.1097/MLR.0b013e31829b1dbd

High Level Expert Group on AI; European Commission. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *ArXiv:1712.09923 [Cs, Stat]*. http://arxiv.org/abs/1712.09923

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, *9*(4). https://doi.org/10.1002/widm.1312

Honneth, A. (1993). *The critique of power: Reflective stages in a critical social theory* (3rd ed). MIT Press.

Honneth, A. (1995). *The fragmented world of the social: Essays in social and political philosophy*. State University of New York Press.

Honneth, A. (2007). *Disrespect: The normative foundations of critical theory*. Polity Press.

Honneth, A. (2009). *Pathologies of reason: On the legacy of critical theory*. Columbia University Press.

Honneth, A. (2012). *The I in we: Studies in the theory of recognition*. Polity Press.

Horng, S., Sontag, D. A., Halpern, Y., Jernite, Y., Shapiro, N. I., & Nathanson, L. A. (2017). Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, *12*(4), e0174708. https://doi.org/10.1371/journal.pone.0174708

Hu, F., Jiang, J., & Yin, P. (2020). Prediction of potential commercially inhibitors against SARS-CoV-2 by multi-task deep model. *ArXiv:2003.00728 [q-Bio]*. http://arxiv.org/abs/2003.00728

Hu, Z., Ge, Q., Li, S., Jin, L., & Xiong, M. (n.d.). *Artificial intelligence forecasting of COVID-19 in China. arXiv preprint:2002.07112*. https://arxiv.org/abs/2002.07112

Huang, Y. (2004). The SAR epidemic and its aftermath in China: A political perspective. In S. Knobler, S. Lemon, A. Mack, L. Sivitz, & K. Oberholtzer (Eds.), *Learning from SARS: Preparing for the next disease outbreak*[Workshop Summary] (pp. 116–136). National Academies Press (US). http://www.ncbi.nlm.nih.gov/books/NBK92462/

Ienca, M., & Vayena, E. (2020). On the responsible use of digital data to tackle the COVID-19 pandemic. *Nature Medicine*, *26*(4), 463–464. https://doi.org/10.1038/s41591-020-0832-5

Inan, O. T., Baran Pouyan, M., Javaid, A. Q., Dowling, S., Etemadi, M., Dorier, A., Heller, J. A., Bicen, A. O., Roy, S., De Marco, T., & Klein, L. (2018). Novel wearable seismocardiography and machine learning algorithms can assess clinical status of heart failure patients. *Circulation: Heart Failure*, *11*(1). https://doi.org/10.1161/CIRCHEARTFAILURE.117.004313

Information Commissioner's Office, & The Alan Turing Institute. (2020). *ICO and the Turing consultation on explaining AI decisions guidance*. https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/

Institute Of Electrical And Electronics Engineers-USA. (2017). *IEEE-USA position statement: Artificial Intelligence research, development and regulation*. https://ieeeusa.org/wp-content/uploads/2017/10/AI0217.pdf

Institute Of Electrical And Electronics Engineers. (2018). *The IEEE Global Initiative on ethics of autonomous and intelligent systems.* Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

Jasanoff, S. (2006). Transparency in public science: Purpose, reasons, limits. *Law and Contemporary Problems*, *69*(3), 21–45.

Jasanoff, S. (2012). *Science and public reason*. Routledge.

Jasanoff, S. (2016). *The ethics of invention: Technology and the human future* (1st ed.). W.W. Norton & Company.

Jia, X., Ren, L., & Cai, J. (2020). Clinical implementation of AI technologies will require interpretable AI models. *Medical Physics*, *47*(1), 1–4. https://doi.org/10.1002/mp.13891

Jirotka, M., Grimpe, B., Stahl, B., Eden, G., & Hartswood, M. (2017). Responsible research and innovation in the digital age. *Communications of the ACM*, *60*(5), 62–68. https://doi.org/10.1145/3064940

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jovanovic, M., Radovanovic, S., Vukicevic, M., Van Poucke, S., & Delibasic, B. (2016). Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artificial Intelligence in Medicine*, *72*, 12–21. https://doi.org/10.1016/j.artmed.2016.07.003

Jugov, T., & Ypi, L. (2019). Structural injustice, epistemic opacity, and the responsibilities of the oppressed. *Journal of Social Philosophy*, *50*(1), 7–27. https://doi.org/10.1111/josp.12268

Jumper, J., Tunyasuvunakool, K., Kohli, P., Hassabis, D., & AlphaFold Team. (2020, April 8). *Computational predictions of protein structures associated with COVID-19*. Deepmind. https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, *18*(1), 24. https://doi.org/10.1186/s12874-018-0482-1

Kelly, T. (1998) *Arguing Safety: A Systematic Approach to Managing Safety Cases. Doctoral Thesis*. University of York: Department of Computer Science.

Kelly, T., & McDermid, J. (1998). Safety case patterns-reusing successful arguments. *IEEE Colloquium on Understanding Patterns and Their Application to System Engineering, London*.

Kelly, T. (2003). *A Systematic Approach to Safety Case Management*. SAE International.

Kelly, T., & Weaver, R. (2004). The goal structuring notation–a safety argument notation. In *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*.

Khalifa, M., Magrabi, F., & Gallego, B. (2019). Developing a framework for evidence-based grading and assessment of predictive tools for clinical decision support. *BMC Medical Informatics and Decision Making*, *19*(1), Article 207. https://doi.org/10.1186/s12911-019-0940-7

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2288–2296. https://people.csail.mit.edu/beenkim/papers/KIM2016NIPS_MMD.pdf

Klasnja, P., Consolvo, S., Choudhury, T., Beckwith, R., & Hightower, J. (2009). Exploring privacy concerns about personal sensing. In H. Tokuda, M. Beigl, A. Friday, A. J. B. Brush, & Y. Tobe (Eds.), *Pervasive computing* (Vol. 5538, pp. 176–183). Springer. https://doi.org/10.1007/978-3-642-01516-8_13

Klinenberg, E. (2003). *Heat wave: A social autopsy of disaster in Chicago*. University of Chicago Press.

Koh, P. W., Pierson, E., & Kundaje, A. (2016). *Denoising genome-wide Histone ChIP-seq with convolutional neural networks* [Preprint]. Bioinformatics. https://doi.org/10.1101/052118

Kohler, K., & Scharte, B. (2020). *Integrating AI into civil protection* (No. 260; CSS Analyses in Security Policy). Center for Security Studies ETH Zurich. https://css.ethz.ch/en/services/digital-library/publications/publication.html/1ec7a3ab-c550-4689-807d-806e8cbd8365

Kristal, T., Cohen, Y., & Navot, E. (2018). Benefit inequality among American workers by gender, race, and ethnicity, 1982–2015. *Sociological Science*, *5*, 461–488. https://doi.org/10.15195/v5.a20

Kruse, C. S., Goswamy, R., Raval, Y., & Marawi, S. (2016). Challenges and opportunities of big data in health care: A systematic review. *JMIR Medical Informatics*, *4*(4), Article e38. https://doi.org/10.2196/medinform.5359

Kuhse, H., & Singer, P. (Eds.). (2009). *A companion to bioethics* (2nd ed). Wiley.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1675–1684. https://doi.org/10.1145/2939672.2939874

Latour, B. (1993). *We have never been modern*. Harvard University Press.

Lee, V. J., Chiew, C. J., & Khong, W. X. (2020). Interrupting transmission of COVID-19: Lessons from containment efforts in Singapore. *Journal of Travel Medicine*, *27*(3), Article taaa039. https://doi.org/10.1093/jtm/taaa039

Lee, Y. (2020, March 27). Taiwan's carrot-and-stick approach to virus fight wins praise, but strains showing. Reuters. https://uk.reuters.com/article/us-health-coronavirus-taiwan-quarantine/taiwans-carrot-and-stick-approach-to-virus-fight-wins-praise-but-strains-showing-idUKKBN21E0EE

Lehr, D., & Ohm, P. (2017). Playing with the data: What legal scholars should learn about machine learning. *University of California, Davis*, *51*, 653–717.

Leonelli, S. (2013). Why the current insistence on open access to scientific data? Big data, knowledge production, and the political economy of contemporary biology. *Bulletin of Science, Technology & Society*, *33*(1–2), 6–11. https://doi.org/10.1177/0270467613496768

Leonelli, S. (2019). Data—From objects to assets. *Nature*, *574*(7778), 317–320. https://doi.org/10.1038/d41586-019-03062-w

Leslie, D. (2016). Machine intelligence and the ethical grammar of computability. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 63–78). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_5

Leslie, D. (2019a). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute. https://doi.org/10.5281/ZENODO.3240529

Leslie, D. (2019b). Raging robots, hapless humans: The AI dystopia. Nature, 574(7776), 32–33. https://doi.org/10.1038/d41586-019-02939-0

Leslie, D. (2020). Isaac Asimov: Centenary of the great explainer. *Nature*, *577*(7792), 614–616. https://doi.org/10.1038/d41586-020-00176-4

Leslie, D., & Cowls, J. (2020, January 6). *For trustworthy AI, explaining "why" is essential*. TechUK. https://www.techuk.org/insights/opinions/item/16523-for-trustworthy-ai-explaining-why-is-essential

Leslie, D., Holmes, L., Hitrova, C., & Ott, E. (2020). *Ethics review of machine learning in children's social care* (p. 74). What Works for Children's Social Care. https://whatworks-csc.org.uk/wp-ontent/uploads/WWCSC_Ethics_of_Machine_Learning_in_CSC_Jan2020.pdf

Li, O., Liu, H., Chen, C., & Rudin, C. (2018). *Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions*. The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17082/16552

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2017). Learning to diagnose with LSTM recurrent neural networks. *ArXiv:1511.03677 [Cs]*. http://arxiv.org/abs/1511.03677

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014). Early diagnosis of Alzheimer's disease with deep learning. *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, 1015–1018. https://doi.org/10.1109/ISBI.2014.6868045

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '12*, 150. https://doi.org/10.1145/2339530.2339556

Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., & Yang, Y. (2014). Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, *35*(28), 2040–2046. https://doi.org/10.1002/jcc.23718

MacIntyre, C. R., & Travaglia, J. F. (2015). Heightened vulnerability, reduced oversight, and ethical breaches on the internet in the West African Ebola epidemic. *The American Journal of Bioethics*, *15*(4), 65–68. https://doi.org/10.1080/15265161.2015.1010017

Maher, N. A., Senders, J. T., Hulsbergen, A. F. C., Lamba, N., Parker, M., Onnela, J.-P., Bredenoord, A. L., Smith, T. R., & Broekman, M. L. D. (2019). Passive data collection and use in healthcare: A systematic review of ethical issues. *International Journal of Medical Informatics*, *129*, 242–247. https://doi.org/10.1016/j.ijmedinf.2019.06.015

Mallett, S., Royston, P., Dutton, S., Waters, R., & Altman, D. G. (2010). Reporting methods in studies developing prognostic models in cancer: A review. *BMC Medicine*, *8*(1), Article 20. https://doi.org/10.1186/1741-7015-8-20

Manrai, A. K., Funke, B. H., Rehm, H. L., Olesen, M. S., Maron, B. A., Szolovits, P., Margulies, D. M., Loscalzo, J., & Kohane, I. S. (2016). Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine, 375*(7), 655–665. https://doi.org/10.1056/NEJMsa1507092

Marks, M. (2020). *Emergent medical data: Health Information inferred by artificial intelligence*. *U.C. Irvine Law Review* (2021, Forthcoming). SSRN: https://ssrn.com/abstract=3554118

May, L., & Delston, J. (Eds.). (2016). *Applied ethics: A multicultural approach* (6th. ed.). Routledge, Taylor & Francis Group.

McNutt, M. (2014). Reproducibility. *Science*, *343*(6168), 229–229. https://doi.org/10.1126/science.1250475

Medawar, P. (1967). *The art of the soluble*. Metheun.

Medicines & Healthcare products Regulatory Agency (MHRA). (2018). *'GXP' Data Integrity Guidance and Definitions* (p. 21). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687246/MHRA_GxP_data_integrity_guide_March_edited_Final.pdf

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ArXiv:1908.09635 [Cs]*. http://arxiv.org/abs/1908.09635

Mejova, Y., & Kalimeri, K. (2020). Advertisers jump on coronavirus bandwagon: Politics, news, and business. *ArXiv:2003.00923 [Cs]*. http://arxiv.org/abs/2003.00923

Mendoza, G., Levine, R., Kibuka, T., & Okoko, L. (2014). *MHealth compendium* (African Strategies for Health, Management Science for Health Volume Four). USAID. http://www.africanstrategies4health.org/uploads/1/3/5/3/13538666/usaid_mHealth_compendium_vol._4_final.pdf

Merson, L., Phong, T. V., Nhan, L. N. T., Dung, N. T., Ngan, T. T. D., Kinh, N. V., Parker, M., & Bull, S. (2015). Trust, respect, and reciprocity: Informing culturally appropriate data-sharing practice in Vietnam. *Journal of Empirical Research on Human Research Ethics*, *10*(3), 251–263. https://doi.org/10.1177/1556264615592387

Metsky, H. C., Freije, C. A., Kosoko-Thoroddsen, T.-S. F., Sabeti, P. C., & Myhrvold, C. (2020). *CRISPR-based surveillance for COVID-19 using genomically-comprehensive machine learning design* [Preprint]. Genomics. https://doi.org/10.1101/2020.02.26.967026

Mill, J. S. (1859/2006). *On liberty and the subjection of a women*. Penguin.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Minsky, M. L. (2015). *Semantic information processing*. The MIT Press.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, *19*(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mirowski, P. (2002). *Machine dreams: Economics becomes a cyborg science*. Cambridge University Press.

Mirowski, P. (2011). *Science-mart: Privatizing American science*. Harvard University Press.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–288. https://doi.org/10.1145/3287560.3287574

Modjarrad, K., Moorthy, V. S., Millett, P., Gsell, P.-S., Roth, C., & Kieny, M.-P. (2016). Developing global norms for sharing data and results during public health emergencies. *PLOS Medicine*, *13*(1), Article e1001935. https://doi.org/10.1371/journal.pmed.1001935

Molloy, J. C. (2011). The Open Knowledge Foundation: Open data means better science. *PLoS Biology*, *9*(12), Article e1001195. https://doi.org/10.1371/journal.pbio.1001195

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. https://christophm.github.io/interpretable-ml-book/

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, *162*(1), W1. https://doi.org/10.7326/M14-0698

Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, *98*(9), 691–698. https://doi.org/10.1136/heartjnl-2011-301247

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2019). *The debate on the ethics of AI in health care: A reconstruction and critical review*. *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3486518

Mozur, P., Zhong, R., & Krolik, A. (2020, March 1). In coronavirus fight, China gives citizens a color code, with red flags. *The New York Times*. https://www.nytimes.com/2020/03/01/business/china-coronavirus-surveillance.html

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, *107*(5), 476–480. https://doi.org/10.1257/aer.p20171084

Nanni, M., Andrienko, G., Barabási, A.-L., Boldrini, C., Bonchi, F., Cattuto, C., Chiaromonte, F., Comandé, G., Conti, M., Coté, M., Dignum, F., Dignum, V., Domingo-Ferrer, J., Ferragina, P., Giannotti, F., Guidotti, R., Helbing, D., Kaski, K., Kertesz, J., … Vespignani, A. (2020). Give more data, awareness and control to individual citizens, and they will help COVID-19 containment. *ArXiv:2004.05222 [Cs]*. http://arxiv.org/abs/2004.05222

Nathanson, A. D., Srikantha, A., Zeidler, D., Wojek, C., Tate, M. L. K., Team, M., & Ag, C. Z. (2019). *Towards cellular epidemiology of degenerative diseases using multibeam SEM and machine learning approaches* (No. 0085; ORS 2019 Annual Meeting Paper). Orthopaedic Research Society. https://www.ors.org/Transactions/65/0085.pdf

National Academies of Sciences, Engineering, and Medicine (U.S.) (Eds.). (2018). *Open science by design: Realizing a vision for 21st century research*. National Academies Press.

National Academies of Sciences, Engineering, and Medicine (U.S.) (Eds.). (2019). *Reproducibility and replicability in science*. National Academies Press.

Nauck, D., & Kruse, R. (1999). Obtaining interpretable fuzzy classification rules from medical data. *Artificial Intelligence in Medicine*, *16*(2), 149–169. https://doi.org/10.1016/S0933-3657(98)00070-0

Ng, J. H., Ye, F., Ward, L. M., Haffer, S. C. "Chris," & Scholle, S. H. (2017). Data on race, ethnicity, and language largely incomplete for managed care plan members. *Health Affairs*, *36*(3), 548–552. https://doi.org/10.1377/hlthaff.2016.1044

Niiler, E. (2020, January 25). An AI epidemiologist sent the first alerts of the coronavirus. *Wired*. https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings/

Nissenbaum, H. F. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford Law Books.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nuffield Council on Bioethics. (2020). *Research in global health emergencies: Ethical issues*. https://www.nuffieldbioethics.org/publications/research-in-global-health-emergencies

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Grady, W. (2015). Processing determinism: Processing determinism. *Language Learning*, *65*(1), 6–32. https://doi.org/10.1111/lang.12091

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA LAW REVIEW*, *57*, 1701–1777.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books.

O'Neil, C. (2020, April 15). *The COVID-19 tracking app won't work—Bloomberg*. Bloomberg Opinion. https://www.bloomberg.com/opinion/articles/2020-04-15/the-covid-19-tracking-app-won-t-work

Organisation for Economic Co-operation and Development. (2019a). *OECD principles on AI*. https://www.oecd.org/going-digital/ai/principles/

Organisation for Economic Co-operation and Development. (2019b). *Artificial intelligence in society*. https://doi.org/10.1787/eedfee77-en

Owen, R., Macnaghten, P., & Stilgoe, J. (2012). Responsible research and innovation: From science in society to science for society, with society. *Science and Public Policy*, *39*(6), 751–760. https://doi.org/10.1093/scipol/scs093

Owen, R. (2014). The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation. *Journal of Responsible Innovation*, *1*(1), 113–117. https://doi.org/10.1080/23299460.2014.882065

Pan-European Privacy-Preserving Proximity Tracing. (2020). *Overview: How We Preserve Privacy and Maintain Security*. https://www.pepp-pt.org/content

Panch, T., Szolovits, P., & Atun, R. (2018). Artificial intelligence, machine learning and health systems. *Journal of Global Health*, *8*(2), Article 020303. https://doi.org/10.7189/jogh.08.020303

Park, D. H., Hendricks, L. A., Akata, Z., Schiele, B., Darrell, T., & Rohrbach, M. (2017). Attentive explanations: Justifying decisions and pointing to the evidence. *ArXiv:1612.04757 [Cs]*. http://arxiv.org/abs/1612.04757

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 39–48. https://doi.org/10.1145/3287560.3287567

Peirce, C. S. (1868). Some consequences of four incapacities. *Journal of Speculative Philosophy*, *2*, 140–157.

Peirce, C. S. (1877). The fixation of belief. *Popular Science Monthly*, *12*, 1–15.

Pfohl, S., Duan, T., Ding, D. Y., & Shah, N. H. (2019). Counterfactual reasoning for fair clinical risk prediction. *ArXiv:1907.06260 [Cs, Stat]*. http://arxiv.org/abs/1907.06260

Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., & Shah, N. H. (2019). Creating fair models of atherosclerotic cardiovascular disease risk. *ArXiv:1809.04663 [Cs, Stat]*. http://arxiv.org/abs/1809.04663

Picardi, C., Hawkins, R., Paterson, C., & Habli, I. (2019). A pattern for arguing the assurance of machine learning in medical diagnosis systems. In A. Romanovsky, E. Troubitsyna, & F. Bitsch (Eds.), *Computer safety, reliability, and security* (Vol. 11698, pp. 165–179). Springer International Publishing. https://doi.org/10.1007/978-3-030-26601-1_12

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature, 473*(7347), 285–285. https://doi.org/10.1038/473285a

Popejoy, A. B., Ritter, D. I., Crooks, K., Currey, E., Fullerton, S. M., Hindorff, L. A., Koenig, B., Ramos, E. M., Sorokin, E. P., Wand, H., Wright, M. W., Zou, J., Gignoux, C. R., Bonham, V. L., Plon, S. E., Bustamante, C. D., & Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG). (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (REA) in genomics. *Human Mutation*, *39*(11), 1713–1720. https://doi.org/10.1002/humu.23644

Pourhomayoun, M., & Shakibi, M. (2020). *Predicting mortality risk in patients with COVID-19 Using artificial intelligence to help medical decision-making* [Preprint]. Health Informatics. https://doi.org/10.1101/2020.03.30.20047308

Prasad, A., & Kotz, D. (2017). ENACT: Encounter-based Architecture for Contact Tracing. *Proceedings of the 4th International on Workshop on Physical Analytics - WPA '17*, 37–42. https://doi.org/10.1145/3092305.3092310

Qi, X., Jiang, Z., Yu, Q., Shao, C., Zhang, H., Yue, H., Ma, B., Wang, Y., Liu, C., Meng, X., Huang, S., Wang, J., Xu, D., Lei, J., Xie, G., Huang, H., Yang, J., Ji, J., Pan, H., … Ju, S. (2020). *Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.02.29.20029603

Raskar, R., Schunemann, I., Barbar, R., Vilcans, K., Gray, J., Vepakomma, P., Kapa, S., Nuzzo, A., Gupta, R., Berke, A., Greenwood, D., Keegan, C., Kanaparti, S., Beaudry, R., Stansbury, D., Arcila, B. B., Kanaparti, R., Pamplona, V., Benedetti, F. M., … Werner, J. (2020). Apps gone rogue: Maintaining personal privacy in an epidemic. *ArXiv:2003.08567 [Cs]*. http://arxiv.org/abs/2003.08567

Reddy, E., Kumar, S., Rollings, N., & Chandra, R. (2015). Mobile Application for dengue fever monitoring and tracking via GPS: Case study for Fiji. *ArXiv:1503.00814 [Cs]*. http://arxiv.org/abs/1503.00814

Reichert, L., Brack, S., & Scheuermann, B. (2020). *Privacy-preserving contact tracing of COVID-19 Patients*. International Association for Cryptologic Research. https://eprint.iacr.org/2020/375.pdf

Riley, R. D., Ensor, J., Snell, K. I. E., Harrell, F. E., Martin, G. P., Reitsma, J. B., Moons, K. G. M., Collins, G., & van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, Article m441. https://doi.org/10.1136/bmj.m441

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2019). The Chinese approach to artificial intelligence: An analysis of policy and regulation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3469784

Rogers, M. L. (2009). *The undiscovered Dewey: Religion, morality, and the ethos of democracy*. Columbia University Press.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638. https://doi.org/10.1017/S0269888913000039

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust1. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Royce, J. (1908/1995). *The philosophy of loyalty*. Vanderbilt University Press.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *ArXiv:1811.10154 [Cs, Stat]*. http://arxiv.org/abs/1811.10154

Rudin, C., & Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, *48*(5), 449–466. https://doi.org/10.1287/inte.2018.0957

Ustun, B., & Rudin, C. (2019). Learning Optimized Risk Scores. *ArXiv:1610.00168 [Math, Stat]*. http://arxiv.org/abs/1610.00168

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control* (1st ed.). Allen Lane.

Sacks, J. A., Zehe, E., Redick, C., Bah, A., Cowger, K., Camara, M., Diallo, A., Gigo, A. N. I., Dhillon, R. S., & Liu, A. (2015). Introduction of mobile health tools to support Ebola surveillance and contact tracing in Guinea. *Global Health: Science and Practice*, *3*(4), 646–659. https://doi.org/10.9745/GHSP-D-15-00207

Sadilek, A., Caty, S., DiPrete, L., Mansour, R., Schenk, T., Bergtholdt, M., Jha, A., Ramaswami, P., & Gabrilovich, E. (2018). Machine-learned epidemiology: Real-time detection of foodborne illness at scale. *Npj Digital Medicine*, *1*(1), Article 36. https://doi.org/10.1038/s41746-018-0045-1

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, *6*(1), Article 205395171882054. https://doi.org/10.1177/2053951718820549

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3126971

Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard University Press. (Original work published 1956)

Sellars, W. (2007). *In the space of reasons: Selected essays of Wilfrid Sellars*. Harvard University Press.

Sen, A. (2011). *The idea of justice* (pbk. ed). Belknap Press of Harvard University Press.

Sengers, P., Boehner, K., David, S., & Kaye, J. "Jofish." (2005). Reflective design. *Proceedings of the 4th Decennial Conference on Critical Computing between Sense and Sensibility - CC '05*, 49. https://doi.org/10.1145/1094562.1094569

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shahabi, C., Fan, L., Nocera, L., Xiong, L., & Li, M. (2015). Privacy-preserving inference of social relationships from location data: A vision paper. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '15*, 1–4. https://doi.org/10.1145/2820783.2820880

Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., Shen, D., & Shi, Y. (2020). Lung infection quantification of COVID-19 in CT images with deep learning. *ArXiv:2003.04655 [Cs, Eess, q-Bio]*. http://arxiv.org/abs/2003.04655

Shapin, S. (1996). *The scientific revolution*. University of Chicago Press.

Shapin, S. (2007). Science and the modern world. In *The Handbook of Science and Technology Studies* (3rd ed.). MIT Press.

Shapin, S. (2010). *Never pure: Historical studies of science as if it was produced by people with bodies, situated in time, space, culture, and society, and struggling for credibility and authority*. Johns Hopkins University Press.

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air-pump: Hobbes, boyle, and the experimental life*. Princeton University Press.

Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *JAMA, 320*(21), 2199–2200. https://doi.org/10.1001/jama.2018.17163

Shrum, W. (2005). Reagency of the internet, or, how I became a guest for science. *Social Studies of Science*, *35*(5), 723–754. https://doi.org/10.1177/0306312705052106

Singer, P. (1979). *Practical ethics*. Cambridge University Press.

Singer, P. (Ed.). (1986). *Applied ethics*. Oxford University Press.

Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. A. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007

Siracusa Principles on the Limitation and Derogation of Provisions in the International Covenant on Civil and Political Rights, Annex, UN Doc E/CN.4/1984/4 (1984). https://www.refworld.org/docid/4672bc122.html

Smith, R. J., Grande, D., & Merchant, R. M. (2016). Transforming scientific inquiry: Tapping Into digital data by building a culture of transparency and consent. *Academic Medicine*, *91*(4), 469–472. https://doi.org/10.1097/ACM.0000000000001022

Spencer, K., Sanders, C., Whitley, E. A., Lund, D., Kaye, J., & Dixon, W. G. (2016). Patient perspectives on sharing anonymized personal health data using a digital system for dynamic consent and research feedback: A qualitative study. *Journal of Medical Internet Research, 18*(4), Article e66. https://doi.org/10.2196/jmir.5011

Srnicek, N. (2017). *Platform capitalism*. Polity.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems*, *86*, 152–161. https://doi.org/10.1016/j.robot.2016.08.018

Star, S. L. (1999). The ethnography of infrastructure. *American Behavioral Scientist*, *43*(3), 377–391. https://doi.org/10.1177/00027649921955326

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., Zhang, J., Chan, L., & Cao, R. (2019). Survey of machine learning techniques in drug discovery. *Current Drug Metabolism*, *20*(3), 185–193. https://doi.org/10.2174/1389200219666180820112457

Stockdale, J., Cassell, J., & Ford, E. (2019). "Giving something back": A systematic review and ethical enquiry into public views on the use of patient data for research in the United Kingdom and the Republic of Ireland. *Wellcome Open Research*, *3*, 6. https://doi.org/10.12688/wellcomeopenres.13531.2

Suresh, H., & Guttag, J. V. (2020). A Framework for Understanding Unintended Consequences of Machine Learning. *ArXiv:1901.10002 [Cs, Stat]*. http://arxiv.org/abs/1901.1000

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (No. 3; Data Privacy Working Paper, p. 34). Carnegie Mellon University.

TCN Coalition. (2020). *TCN protocol*. https://github.com/TCNCoalition/TCN

Taylor, C. (1980). *The explanation of behaviour* (pbk. ed.). Routledge & Kegan Paul [u.a.]. (Original work published 1964)

Taylor, C. (1989). *Sources of the self: The making of the modern identity*. Harvard University Press.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE*, *6*(6), Article e21101. https://doi.org/10.1371/journal.pone.0021101

Todrys, K. W., Howe, E., & Amon, J. J. (2013). Failing Siracusa: Governments' obligations to find the least restrictive options for tuberculosis control. *Public Health Action*, *3*(1), 7–10(4). http://dx.doi.org/10.5588/pha.12.0094

Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. *ArXiv:1905.05134 [Cs, Stat]*. http://arxiv.org/abs/1905.05134

Troncoso, C., Payer, M., Salathé, M., Larus, J., Lueks, D. W., Stadler, T., Pyrgelis, D. A., Antonioli, D., Barman, L., Chatel, S., Paterson, K., Capkun, S., Basin, D., Jackson, D., Leuven, K., Preneel, B., Smart, N., Singelee, D. D., Abidin, D. A., … Cremers, C. (2020). *Decentralized privacy-preserving proximity tracing*. GitHub. https://github.com/DP-3T/documents/blob/master/DP3T%20White%20Paper.pdf

University of Montreal. (2017). *Declaration of Montréal for a responsible development of AI*. Respaideclaration. https://www.montrealdeclaration-responsibleai.com

Uslaner, E. M. (2002). *The moral foundations of trust*. Cambridge University Press.

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349–391. https://doi.org/10.1007/s10994-015-5528-6

Van Bavel, J. J., Baicker, K., Boggio, P., Capraro, V., Cichocka, A., Crockett, M., Cikara, M., Crum, A., Douglas, K., Druckman, J., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J., Gelfand, Mi., Han, S., Haslam, S. A., Jetten, J., … Willer, R. (2020). *Using social and behavioural science to support COVID-19 pandemic response* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/y38m9

van der Aalst, W. M. P., Bichler, M., & Heinzl, A. (2017). Responsible data science. *Business & Information Systems Engineering*, *59*(5), 311–313. https://doi.org/10.1007/s12599-017-0487-z

van der Schaar, M., Humphrey, J., Alaa, A., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., & Ercole, A. (2020). *How artificial intelligence and machine learning can help healthcare systems respond to COVID-19*. https://www.vanderschaar-lab.com/NewWebsite/covid-19/post1/paper.pdf

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), Article e1002689. https://doi.org/10.1371/journal.pmed.1002689

Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-019-04051-w

Vellido, A., Martın-Guerrero, J. D., & Lisboa, P. J. G. (2012). Making machine learning models interpretable. *In ESANN* (Vol. 12, pp. 163–172). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.431.5382
&rep=rep1&type=pdf

Verheij, R. A., Curcin, V., Delaney, B. C., & McGilchrist, M. M. (2018). Possible sources of bias in primary care electronic health record data use and reuse. *Journal of Medical Internet Research*, *20*(5), Article e185. https://doi.org/10.2196/jmir.9134

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness - FairWare ’18*, 1–7. https://doi.org/10.1145/3194770.3194776

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, Article l6927. https://doi.org/10.1136/bmj.l6927

Von Neumann, J. (1955). Can we survive technology? *Fortune*. http://geosci.uchicago.edu/~kite/doc/von_Neumann_1955.pdf

von Schomberg, R. (2011). *Towards responsible research and innovation in the information and communication technologies and security technologies fields*. Directorate-General for Research and Innovation (European Commission). https://op.europa.eu/en/publication-detail/-/publication/60153e8a-0fe9-4911-a7f4-1b530967ef10/language-en

Von Schomberg, R. (2013). A vision of responsible research and innovation. In R. Owen, J. R. Bessant, & M. Heintz (Eds.), *Responsible innovation* (pp. 51–74). Wiley.

Von Schomberg, R. (2019, January). *Why responsible innovation*. Conference FWO. https://www.researchgate.net/publication/330338135_Why_Responsible_Innovation

Von Wright, G. H. (2004). *Explanation and understanding*. Cornell University Press. (Original work published 1971)

Wainberg, M., Merico, D., Delong, A., & Frey, B. J. (2018). Deep learning in biomedicine. *Nature Biotechnology*, *36*(9), 829–838. https://doi.org/10.1038/nbt.4233

Wang, C. J., Ng, C. Y., & Brook, R. H. (2020). Response to COVID-19 in Taiwan: Big data analytics, new technology, and proactive testing. *JAMA*, *323*(14), 1341–1342. https://doi.org/10.1001/jama.2020.3151

Wang, F., & Rudin, C. (2015). Falling rule lists. *ArXiv:1411.5899 [Cs]*. http://arxiv.org/abs/1411.5899

Wang, R. Y., Storey, V. C., & Firth, C. P. (1995). A framework for analysis of data quality research. *IEEE Transactions on Knowledge and Data Engineering*, *7*(4), 623–640. https://doi.org/10.1109/69.404034

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., & Xu, B. (2020). *A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)* [Preprint]. Infectious Diseases (except HIV/AIDS). https://doi.org/10.1101/2020.02.14.20023028

Wang, Z., & Tang, K. (2020). Combating COVID-19: Health equity matters. *Nature Medicine*, *26*(4), 458–458. https://doi.org/10.1038/s41591-020-0823-6

Weiss, D., Rydland, H. T., Øversveen, E., Jensen, M. R., Solhaug, S., & Krokstad, S. (2018). Innovative technologies and social inequalities in health: A scoping review of the literature. *PLOS ONE*, *13*(4), Article e0195447. https://doi.org/10.1371/journal.pone.0195447

Whitlock, M. C. (2011). Data archiving in ecology and evolution: Best practices. *Trends in Ecology & Evolution*, *26*(2), 61–65. https://doi.org/10.1016/j.tree.2010.11.006

Wiemken, T. L., Carrico, R. M., Furmanek, S. P., Guinn, B. E., Mattingly, W. A., Peyrani, P., & Ramirez, J. A. (2020). Socioeconomic position and the incidence, severity, and clinical outcomes of hospitalized patients with community-acquired pneumonia. *Public Health Reports*, Article 003335492091271. https://doi.org/10.1177/0033354920912717

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2016). Interpretable recurrent neural networks using sequential sparse recovery. *ArXiv:1611.07252 [Cs, Stat]*. http://arxiv.org/abs/1611.07252

World Health Organization. (2005). *International health regulations (2005)* (3rd ed.). World Health Organization.

World Health Organization. (2011). *mHealth: New horizons for health through mobile technologies* (Vol. 3; Global Observatory for EHealth Series). World Health Organization.

World Health Organization. (2015, September). *Developing global norms for sharing data and results during public health emergencies*. World Health Organization. https://www.who.int/medicines/ebola-treatment/blueprint_phe_data-share-results/en/

World Health Organization. (2016a). *Annex 5 guidance on good data and record management practices*. https://www.who.int/medicines/publications/pharmprep/WHO_TRS_996_annex05.pdf

World Health Organization. (2016b). *Guidance for managing ethical issues in infectious disease outbreaks*. World Health Organization.

World Health Organization, & Council for International Organizations of Medical Sciences. (2017). *International ethical guidelines for health-related research involving humans*. CIOMS.

Wright, K., Parker, M., & The Nuffield Council on Bioethics Working Group (2020). In emergencies, health research must go beyond public engagement toward a true partnership with those affected. *Nature Medicine*, *26*(3), 308–309. https://doi.org/10.1038/s41591-020-0758-y

Wright, N. (2019a, October 11). *How artificial intelligence will reshape the global order*. https://www.foreignaffairs.com/articles/world/2018-07-10/how-artificial-intelligence-will-reshape-global-order

Wright, N. (2019b). *Artificial intelligence, China, Russia, and the global order: Technological, political, global, and creative perspectives*. Air University Press.

Wynants, L., Van Calster, B., Bonten, M. M. J., Collins, G. S., Debray, T. P. A., De Vos, M., Haller, M. C., Heinze, G., Moons, K. G. M., Riley, R. D., Schuit, E., Smits, L. J. M., Snell, K. I. E., Steyerberg, E. W., Wallisch, C., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. *BMJ*, Article m1328. https://doi.org/10.1136/bmj.m1328

Xu, Y., Biswal, S., Deshpande, S. R., Maher, K. O., & Sun, J. (2018). RAIM: Recurrent Attentive and Intensive Model of multimodal patient monitoring data. *ArXiv:1807.08820 [Cs, Stat]*. http://arxiv.org/abs/1807.08820

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jin, L., Zhang, M., Huang, X., Xiao, Y., Cao, H., Chen, Y., Ren, T., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). *A machine learning-based model for survival prediction in patients with severe COVID-19 infection* [Preprint]. Epidemiology. https://doi.org/10.1101/2020.02.27.20028027

Young, I. M. (1990). *Justice and the politics of difference*. Princeton University Press.

Young, I. M. (2009). Structural injustice and the politics of difference. In T. Christiano & J. Christman (Eds.), *Contemporary debates in political philosophy* (pp. 362–383). Wiley-Blackwell. https://doi.org/10.1002/9781444310399.ch20

Young, I. M., & Allen, D. S. (2011). *Justice and the politics of difference* (pbk. ed.). Princeton University Press.

Yu, K.-H., Zhang, C., Berry, G. J., Altman, R. B., Ré, C., Rubin, D. L., & Snyder, M. (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7(1), Article 12474. https://doi.org/10.1038/ncomms12474

Zastrow, M. (2020, March 18). South Korea is reporting intimate details of COVID-19 cases: Has it helped? *Nature*. https://doi.org/10.1038/d41586-020-00740-y

Zeng, Y., Lu, E., & Huangfu, C. (2019, January 27). *Linking artificial intelligence principles*. AAAI Workshop on Artificial Intelligence Safety SafeAI 2019, Honolulu, Hawaii, USA. http://ceur-ws.org/Vol-2301/paper_15.pdf

Zhang, H., Saravanan, K. M., Yang, Y., Hossain, Md. T., Li, J., Ren, X., & Wei, Y. (2020). *Deep learning based drug screening for novel coronavirus 2019-nCov* [Preprint]. Other. https://doi.org/10.20944/preprints202002.0061.v1

Zhang, Q., Wu, Y. N., & Zhu, S.-C. (2018). Interpretable Convolutional neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8827–8836. https://doi.org/10.1109/CVPR.2018.00920

Zhenwei Qiang, C., Yamamichi, M., Hausman, V., Miller, R., & Altman, D. (2012). *Mobile applications for the health sector* [ICT Sector Unit]. World Bank. http://siteresources.worldbank.org/INFORMATIONANDCOMMUNICATIONANDTECHNOLOGIES/Resources/mHealth_report_(Apr_2012).pdf

Zhu, J., Pande, A., Mohapatra, P., & Han, J. J. (2015). Using deep learning for energy expenditure estimation with wearable sensors. *2015 17th International Conference on E-Health Networking, Application & Services (HealthCom)*, 501–506. https://doi.org/10.1109/HealthCom.2015.7454554

Zink, A., & Rose, S. (2020). Fair Regression for health care spending. *Biometrics*, Article biom.13206. https://doi.org/10.1111/biom.13206

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

Footnotes

[1] Correspondence concerning this article may be sent to the author by email: dleslie@turing.ac.uk ↩

[2] In the following, I will use the abbreviation AI/ML to indicate those information-processing systems or algorithmic models that intervene in the human world (directly or through the insights they enable) by carrying out cognitive or perceptual functions previously reserved for human beings. This broadly pragmatic and functionalist definition is meant to be as generally applicable to both deterministic and nondeterministic algorithm-based computing machinery as it is non-metaphysical (IEEE-USA, 2017; Leslie 2019b; Minsky, 1968/2015; Organisation for Economic Co-operation and Development [OECD], 2019a, 2019b). Similarly, I will refer to 'data science' as the broad, interdisciplinary set of approaches and techniques that combine statistics, applied mathematics, data mining, computer programming, and other related fields to gain a better conceptual understanding and practical grasp of the data patterns underlying the empirical world. ↩

[3] Some examples of these contributions include: in diagnostics, detecting diabetic retinopathy (Gulshan et al., 2016), using waveform analysis to identify birthing paths (Fergus et al., 2017), early diagnosis of Alzheimer disease (Liu et al., 2014), detecting lymph node metastases in women with breast cancer (Bejnordi et al., 2017), classifying sepsis in emergency departments (Horng, 2017), using clinical measurements to classify patients in a pediatric ICU (Lipton et al., 2017), classifying skin cancer (Esteva et al., 2017), diagnosing acute coronary syndrome (Berikol, 2016); in prognostics, predicting breast cancer survival (Katzman et al., 2018), predicting heart condition–related hospitalization (Brisimi, 2018), predicting outcomes in colorectal cancer (Bychkov, 2018), predicting outcomes in non–small cell lung cancer (Yu et al., 2016); In genomics, predicting the sequence specificities of DNA- and RNA-binding proteins (Alipanahi et al., 2015), denoising genome-wide histone ChIP-seq (Koh et al., 2016), predicting protein structures from protein sequences (Lyons et al., 2014); in epidemiology, understanding outcomes in community-spread pneumonia (Wiemken, 2020), understanding degenerative diseases (Nathanson, 2019), detecting food-borne illness (Sadilek et al., 2018); in mobile monitoring, diagnosing heart failure through wearable technology monitoring (Inan et al., 2018), and estimating energy expenditure with wearable sensors (Zhu et al., 2015). For good additional landscape views, see Miotto et al. (2018), Panch et al. (2018), Stephenson (2019), and Wainberg et al. (2018). ↩

[4] The problem here was that the designers of the model chose health care costs as the measurable proxy for the target concept of ill health. That the former is an insufficient stand-in for the latter becomes clearer when one considers factors such as (1) the challenges to accessing health care

faced by traditionally disadvantaged subpopulations and (2) the challenge of the reduced trust in medical services experienced by historically maltreated social groups. This directly affects their level of engagement in health care systems. Because of correlations between socioeconomic status and race, Black patients (even the insured) are less likely to run up the same level of health care costs as Whites of greater advantage. These insights about erroneous proxies in Obermeyer et al. (2019) follow on from the earlier work of two of the authors on mismeasurement in health policy applications (Mullainathan & Obermeyer, 2017). ↩

5. This domain-sensitivity is crucial in medical decision support systems: Only when interpretable models are designed with a proper understanding of the missing data mechanisms endemic to the messiness of the clinical environment of concern, can they generate outputs that are appropriately responsive to its complexities and uncertainties (Ghassemi, 2018). ↩

6. Another crucial component of building this sort of high-performance, high-interpretability model is the incorporation of domain knowledge to ensure that expert understanding of clinical conditions and underlying biological mechanisms is both informing feature selection and ultimately supporting the rationale behind the predictions—though a utilization of domain knowledge should not steer knowledge discovery away from exhaustive search of feature importance beyond existing insights (Gilvary et al., 2019; Jovanovic et al., 2016) ↩

7. Impactful contributions of the interpretability culture have included interpretable decision sets (Lakkaraju, 2016), generalized additive models (Lou et al., 2012), supersparse linear integer models (Rudin & Ustun, 2018, 2019), certifiably optimal rule lists (Angelino et al., 2018), falling rule lists (Wang & Rudin, 2015), Boolean decision rules via column generation (Dash et al., 2018), and case-based reasoning (Bichindaritz & Marling, 2006; Bien & Tibshirani, 2011; Kim et al., 2016, adding criticism to prototypes). ↩

8. Many other strategies to design interpretable deep-learning systems have also been investigated. For instance, Bau et al. (2017), Wisdom et al. (2016), and Zhang et al. (2018). ↩

9. A notable initiative in this direction has already been made by the Research Data Alliance (Berman & Crosas, 2020). In the area of health data, the United Kingdom's Health Data Research UK (HDR UK) is also making major strides forward in institutionalizing responsible data sharing as is the Coleridge Initiative. ↩

10. It would be helpful to note that there has been a high degree of critical self-reflection in RRI about limitations in the generalizability and succinctness of its framing of values-based research and discovery. For instance, issues have been raised about its naïve grouping of the classes of

research and innovation, about inexorable definition disagreements regarding action-orienting values, and about the potential for abuse and misuse of public engagement methods (Jirotka et al., 2017). ↩

11. This context-aware and anticipatory approach has been developed in the area of argument-based assurance of safety-critical digital technologies. "Assurance cases" or "safety cases" provide an integrative and process-based platform for ensuring that the properties needed to fulfil the high-level normative goals of a computational system and to mitigate its anticipated risks are translated into design actions and documented as interrelated claims, arguments, and evidence. Consolidated standards for system and software assurance include the ISO/IEC/IEEE 15026 series and the Object Management Group's Structured Assurance Case Metamodel (SACM), and various assurance platforms exist such as Goal Structuring Notation (GSN), the Claims, Arguments and Evidence Notation (CAE), and Dynamic Safety Cases (DSC). For further background, see Ashmore et al. (2019), Bloomfield & Bishop (2010), Bloomfield & Netkachova (2014), Denney et al. (2015), Ge et al. (2012), Health Foundation (2012); Kelly (1998, 2003), Kelly & McDermid (1998), Kelly & Weaver (2004), and Picardi et al. (2019). ↩

12. Crucially, these ethical principles have arisen in both bioethics and human rights regimes as moral claims that have responded directly to tangible, technologically inflicted harms and atrocities. In a significant sense, that is, both traditions emerged out of concerted public acts of resistance against violence done to disempowered or vulnerable people. Whereas human rights has its origins in efforts to redress the well-known barbarisms and genocides of the mid-twentieth century, in the case of bioethics, its emergence tracked the public exposure in the 1960s and 1970s of several atrocities of human experimentation (such as the infamous Tuskegee syphilis experiment), where it was discovered that members of vulnerable or marginalized social groups had been subjected to the injurious effects of institutionally run biomedical experiments without having knowledge of or giving consent to their participation (Kuhse & Singer, 2009; Leslie et al., 2020). While a longer discussion of this is out of the scope of this article, it is notable that the provisional universalism of AI/ML ethics principles is rooted in a kind of moral grammar that underlies acts of resistance against those who have perpetrated social injury (Honneth, 2007). ↩

13. Recently, scholars have been applying 'fair' ML techniques directly to medicine. For example, Zink & Rose (2019) propose new fair ML modeling methods that use constrained and penalized regression to improve health insurance carrier risk adjustment for undercompensated groups; Pfohl, Marafino, et al. (2019) leverage adversarial learning and EHR data to develop a fair ASCD model; Pfohl, Duan, et al. (2019) uses counterfactual reasoning to apply fairness principles to clinical risk prediction at the individual level. ↩

14. Note that, in this Appendix, I am focussing on the constructive, normative dimension of the history of modern science and, for this reason, am leaving aside the empirical aspects of discursive and institutional power, politics, culture, and socioeconomic stratification that inform other (equally important) critical-sociological histories of modern science (for instance, Foucault, 1961/1988, 1966/2007; Latour, 1993; Mirowski, 2002, 2011; Schaffer & Shapin, 1985; Shapin, 1996, 2010). While both normative and critical-sociological perspectives are crucial, I would suggest that it is also vital to resist disentangling them entirely. That is, from the perspective of a critical theory of society, one must endeavor to discover the sources of normativity that inhere prereflectively in concrete social and historical practices per se, and, only in this way, can one then gain the critical leverage needed to discern those distortions and malformations that manifest in both subtle and explicit forms of power, domination, and coercion (Honneth, 1993, 1995, 2009) . From this critical and ethical-practical point of view, one can gain access to historically effective normativity by reconstructing the conditions of possibility of the particular social practices that lie behind human advancement. This does not mean vacating the interrogation of the sociohistorical fields wherein dispersed and concentrated patterns of violence and power inhere but rather taking a performatively consistent approach to the critique of the latter by first clarifying and making explicit what has gone moral-practically awry. It is this last bit that I'm concentrated on here. ↩

15.

Admittedly, the generic idea of a 'scientific method' is overly schematic as has been pointed out by Medawar (1967), Shapin (2007), and others. Likewise, reference to a unifying idea of 'modern science' has been usefully deconstructed by historians and philosophers of science (for example, in Cartwright, 1999; Dupre, 1993; Galison & Stump, 1996). I use these terms as signposts for particular kinds of distinctively postconventional social practices that carry historically sustained normative relevance rather than as descriptors that are applicable in their historical specificity.

[1] Admittedly, the generic idea of a 'scientific method' is overly schematic as has been pointed out by Medawar (1967), Shapin (2007), and others. Likewise, reference to a unifying idea of 'modern science' has been usefully deconstructed by historians and philosophers of science (for example, in Cartwright, 1999; Dupre, 1993; Galison & Stump, 1996). I use these terms as signposts for particular kinds of distinctively postconventional social practices that carry historically sustained normative relevance rather than as descriptors that are applicable in their historical specificity. ↩