Philosophy & Technology (2020) 33:659–684
https://doi.org/10.1007/s13347-020-00405-8

RESEARCH ARTICLE

Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence

Shakir Mohamed[1] · **Marie-Therese Png**[2] · **William Isaac**[1]

Received: 16 January 2020 / Accepted: 26 May 2020 / Published online: 12 July 2020

Abstract

This paper explores the important role of critical science, and in particular of post-colonial and decolonial theories, in understanding and shaping the ongoing advances in artificial intelligence. Artificial intelligence (AI) is viewed as amongst the technological advances that will reshape modern societies and their relations. While the design and deployment of systems that continually adapt holds the promise of far-reaching positive change, they simultaneously pose significant risks, especially to already vulnerable peoples. Values and power are central to this discussion. Decolonial theories use historical hindsight to explain patterns of power that shape our intellectual, political, economic, and social world. By embedding a decolonial critical approach within its technical practice, AI communities can develop foresight and tactics that can better align research and technology development with established ethical principles, centring vulnerable peoples who continue to bear the brunt of negative impacts of innovation and scientific progress. We highlight problematic applications that are instances of coloniality, and using a decolonial lens, submit three tactics that can form a decolonial field of artificial intelligence: creating a critical technical practice of AI, seeking reverse tutelage and reverse pedagogies, and the renewal of affective and political communities. The years ahead will usher in a wave of new scientific breakthroughs and technologies driven by AI research, making it incumbent upon AI communities to strengthen the social contract through ethical foresight and the multiplicity of intellectual perspectives available to us, ultimately supporting future technologies that enable greater well-being, with the goal of beneficence and justice for all.

Keywords Decolonisation · Coloniality · Sociotechnical foresight · Intercultural ethics · Critical technical practice · Artificial intelligence · Affective community

Shakir Mohamed, Marie-Therese Png and William Isaac contributed equally to this work.

✉ Shakir Mohamed
shakir@deepmind.com

Extended author information available on the last page of the article.

1 How Values Shape Scientific Knowledge and Technology

The ongoing advances in artificial intelligence (AI), and innovations in technology more generally, encompass ever-larger aspects of the cultural, economic and political life of modern society. We aim to capture this expanding role and impact of AI by widening the conceptual aperture with which it is understood: dually viewing AI as both object and subject, i.e. viewing AI as technological artefacts and as systems of networks and institutions, respectively.

As an object, advances in AI research[1] over the last two decades—often attributed to a combination of increases in computational power, availability of large amounts of data and advances in learning algorithms (LeCun et al. 2015)—has led to novel applications in a wide range of sectors, including transportation and healthcare, amongst others (Gerrish 2018). While these recent innovations have led to some societal benefits, they have also demonstrated their potential to be abused or misused in ways their designers could not have imagined (O'Neil 2016). As a subject, AI has seen itself elevated from an obscure domain of computer science into technological artefacts embedded within and scrutinised by governments, industry and civil society. These stakeholders play a significant role in shaping the future direction and use of advanced technologies such as AI—whether through the establishment of regulatory and ethical frameworks or the promotion of specific algorithmic architectures[2]—that warrants consideration under a more expansive conceptualisation of the term AI.

As both object and subject, the aims and applications of AI have been brought into question. At the heart of these discussions are questions of values and the power relations in which these values are embedded. What values and norms should we aim to uphold when performing research or deployment of systems based on artificial intelligence? In what ways do failures to account for asymmetrical power dynamics undermine our ability to mitigate identified harms from AI? How do unacknowledged and unquestioned systems of values and power inhibit our ability to assess harms and failures in the future?

1.1 The Evolution of Value and Power Paradigms

The role that values play in the process of generating new knowledge is a perennial question, particularly in the sciences. Philosophers have debated the importance of *epistemic values*, such as internal consistency, falsifiability, generalisability of a particular theory and notions of scientific objectivity (Laudan 1968; Bueter 2015). These

[1]This view of AI as object, and of the term AI throughout, will be used as an umbrella term that includes the field of machine learning. Both machine learning and artificial intelligence are disciplines focused on the science and engineering of intelligent agents or computer programs (Russell S and Norvig P 2016; Boden 2018). While the broader field of AI includes both symbolic (also known as classical AI or GOFAI) and connectionist (e.g. artificial neural networks) research, the field of machine learning can be defined by research on more tractable machine tasks leveraging techniques at the intersection of computer science and statistical inference (Mitchell 2006).

[2]Specific applications of machine learning and AI largely centre on the use of learning *algorithms* that manipulate and transform data into information suitable for the given task.

values shape the veracity of scientific statements, aiming to establish broader ontological or causal claims about the nature of specific systems. Yet, science is a product not only of epistemic values, but also of *contextual values* that reflect moral, societal or personal concerns in the application of scientific knowledge. There is strong consensus that non-epistemic values have a legitimate role in scientific reasoning, particularly in the choice of research projects and the application of scientific results (Elliott and McKaughan 2014; Douglas 2007). This role of contextual values also applies to the work of computing and technology (Nissenbaum 2001; Van de Poel and Kroes 2014)—a recognition established in the broader field of values in technology (Friedman et al. 2013; Sengers et al. 2005; DiSalvo 2012).

Due to repeated instances of unethical research practices within the scientific community—instances like the U.S. Public Health Service Syphilis Study at Tuskegee (Brandt 1978)—concerned scientists, policy-makers and human rights advocates responded by formalising contextual values into ethical frameworks that reoriented power relations between researchers and impacted communities or persons. Efforts such as the Nuremberg Code (Nuremberg Military Tribunals 1947), the Helsinki declaration (WMA General Assembly 1964) and the Belmont Principles (Belmont Commission 1978) collectively led to the establishment of three core ethical values or rights that should serve as a minimum standard for human subject research: respect-for-persons (individual autonomy), beneficence (research designed to maximise societal benefit and minimise individual harm) and justice (research risks must be distributed across society). These principles are viewed as a historical milestone for research ethics, although their violations continue to occur, e.g. the ongoing questions of unethical blood exports during the West Africa ebola epidemic (Freudenthal 2019). These principles are also questioned and subject to many reappraisals, which have highlighted their failures in capturing a range of emerging or novel harms, or an insufficiency in capturing the lived realities of under-represented groups (Shore 2006; Vitak et al. 2016).

The limitations of these value principles become clearer as AI and other advanced technologies become enmeshed within high-stakes spheres of our society. Initial attempts to codify ethical guidelines for AI, e.g. the Asilomar principles (Asilomar Meeting 2017), focused on risks related to lethal autonomous weapons systems and AGI Safety. Though both are critical issues, these guidelines did not recognise that risks in peace and security are first felt by conflict zones in developing countries (Garcia 2019), or engage in a disambiguation of social safety and technical safety. Moreover, they did not contend with the intersection of values and power, whose values are being represented, and the structural inequities that result in an unequal spread of benefits and risk within and across societies.

An example of this nexus between values, power and AI is a recent study by Obermeyer et al. (2019), which revealed that a widely used prediction algorithm for selecting entry into healthcare programs was exhibiting racial bias against African-American patients. The tool was designed to identify patients suitable for enrolment into a "high-risk care management" programme that provides access to enhanced medical resources and support. Unfortunately, large health systems in the USA have emphasised contextual values to "reduce overall costs for the healthcare system

while increasing value" (AMA 2018) or "value for money" (UK National Health Service 2019) on "value for money") when selecting potential vendors for algorithmic screening tools at the expense of other values such as addressing inequities in the health system. As a result, the deployed algorithm relied on the predictive utility of an individual's health expenses (defined as total healthcare expenditure) indirectly leading to the rejection of African-American patients at a higher rate relative to white patients, denying care to patients in need, and exacerbating structural inequities in the US healthcare system (Nelson 2002). As this example shows, the unique manner in which AI algorithms can quickly ingest, perpetuate and legitimise forms of bias and harm represents a step change from previous technologies, warranting prompt reappraisal of these tools to ensure ethical and socially beneficial use.

An additional challenge is that AI can obscure asymmetrical power relations in ways that make it difficult for advocates and concerned developers to meaningfully address during development. As Benjamin (2019) notes, "whereas in a previous era, the intention to deepen racial inequities was more explicit, today coded inequity is perpetuated precisely because those who design and adopt such tools are not thinking carefully about systemic racism". Some scholars such as Floridi et al. (2018) have highlighted that technologies such as AI require an expansion of ethical frameworks, such as the Belmont Principles, to include explicability (explanation and transparency) or non-malfeasance (do no harm). Whittlestone et al. (2019) conversely argue for a move away from enumerating new value criteria, and instead highlight the need to engage more deeply with the tensions that arise between principles and their implementation in practice. Similarly, we argue that the field of AI would benefit from dynamic and robust *foresight* tactics and methodologies grounded in the critical sciences to better identify limitations of a given technology and their prospective ethical and social harms.

1.2 Critical Science as a Sociotechnical Foresight Tool

The critical science approach represents a loosely associated group of disciplines that seek to uncover the underlying cultural assumptions that dominate a field of study and the broader society. Scholarship in this domain (Winner 1980; Nissenbaum 2001; Greene et al. 2019) aims not only to explain sociotechnical phenomena, but to also examine issues of values, culture and power at play between stakeholders and technological artefacts. We use a necessarily broad scope of critical science theories due to the expansive range of applications of AI, but seek to emphasise particularly the role of *post-colonial and decolonial critical theories*. While decolonial studies begins from a platform of historical colonialism, it is deeply entangled with the critical theories of race, feminism, law, queerness and science and technology studies (D'Ignazio and Klein 2020; Feenberg 2017).

The role of values and power as they relate to technology and data has been argued by a multitude of scholars who draw from the decolonial theories, such as Ricaurte (2019), Milan and Treré (2019) and Couldry and Mejias (2019a), as well as established research in post-colonial and decolonial computing, such as those by Irani et al. (2010), Dourish and Mainwaring (2012) and Ali (2016). Such critical perspectives are increasingly used to elucidate potential ethical and social ramifications of AI

and technology generally, with much research now available that exposes concerns of bias and injustice in algorithmic systems (Angwin et al. 2016a; Benjamin 2019; Buolamwini and Gebru 2018; Lum and Isaac 2016; Noble 2018; Eubanks 2018), exploitative or extractive data practices (Gray and Suri 2019), and applications of AI that dispossess the identity and resources of vulnerable populations (Green 2019; Keyes 2018; Hanna et al. 2019; Stark 2019).

In this paper, our aim is to guide the reader through a brief introduction to decolonial theory, and demonstrate how this theoretical framework can serve as a powerful lens of ethical foresight. Technology foresight and foresight methodologies more broadly is a term used to classify efforts by researchers, policy-makers and industry practitioners to understand and anticipate how choices and actions made today can shape or create the future (Coates and at al. 1985). For AI technologies, problematic outcomes in high-stakes domains such as healthcare or criminal justice have demonstrated a clear need for dynamic and robust ethical foresight methodologies. Such methodologies could enable stakeholders to better anticipate and surface blind-spots and limitations, expand the scope of AI's benefits and harms and reveal the relations of power that underlie their deployment. This is needed in order to better align our research and technology development with established and emerging ethical principles and regulation, and to empower vulnerable peoples who, so often, bear the brunt of negative impacts of innovation and scientific progress.

2 Coloniality and Decolonial Theory

Decolonisation refers to the intellectual, political, economic and societal work concerned with the restoration of land and life following the end of historical colonial periods (Ashcroft 2006). Territorial appropriation, exploitation of the natural environment and of human labour, and direct control of social structures are the characteristics of historical colonialism. Colonialism's effects endure in the present, and when these colonial characteristics are identified with present-day activities, we speak of the more general concept of *coloniality* (Quijano 2000; Mignolo 2007; Maldonado-Torres 2007). This section is a brief review of coloniality, its view on systems of power and its manifestation in the digital world.

Coloniality is what survives colonialism (Ndlovu-Gatsheni 2015). Coloniality therefore seeks to explain the continuation of power dynamics between those advantaged and disadvantaged by "the historical processes of dispossession, enslavement, appropriation and extraction [...] central to the emergence of the modern world" (Bhambra et al. 2018). Coloniality names the continuity of established patterns of power between coloniser and colonised—and the contemporary remnants of these relationships—and how that power shapes our understanding of culture, labour, intersubjectivity and knowledge production: what Quijano (2000) refers to as the coloniality of power. For Quijano (2000), the power of coloniality lies in its control over social structures in the four dimensions of authority, economy, gender and sexuality, and knowledge and subjectivity. Similarly, for Maldonado-Torres (2007), coloniality is the reproduction of hierarchies of race, gender and geopolitics, which were invented or instrumentalised as tools of colonial control. For Couldry and

Mejias (2019a), who bring the coloniality of power into the digital present, it is modern data relations—the human relations that when captured as data enables them to become a commodity—that "recreate a colonising form of power".

Consequently, *decolonisation* takes two roles. The first is a territorial decolonisation that is achieved by the dissolution of colonial relations. The second is a structural decolonisation, with which this paper is concerned, that seeks to undo colonial mechanisms of power, economics, language, culture and thinking that shapes contemporary life: interrogating the provenance and legitimacy of dominant forms of knowledge, values, norms and assumptions. Three views clarify this decolonial knowledge landscape.

- A *decentring view* of decolonisation seeks to reject an imitation of the West in all aspects of life, calling for the assertion of unique identities and a re-centring of knowledge on approaches that restore global histories and problems and solutions. For Ngũgĩ wa Thiong'o, this means replacing the English language as the unassailable medium of teaching and discourse (Wa Thiong'o 1992). Discussions on decolonising the curriculum or decolonising the university call for reappraisals of what is considered the foundation of an intellectual discipline by emphasising and recognising the legitimacy of marginalised knowledge (Jansen 2019; Bhambra et al. 2018); calls to decolonise science often invoke this view of decolonisation (Harding 2011).
- An *additive-inclusive view* continues to use existing knowledge, but in ways that recognises explicitly the value of new and alternative approaches, and that supports environments in which new ways of creating knowledge can genuinely flourish. This view is invoked by works that criticise universalism in thinking, and instead advocate for localisation and pluriversalism (Mignolo 2012; Escobar 2011).
- An *engagement view* calls directly for more critical views of science. This view calls on us to examine scientific practice from the margins, to place the needs of marginalised populations at the centre of the design and research process, and to ask where knowledge comes from—who is included and left out, in whose interest is science applied, who is silenced and what unacknowledged assumptions might be at play (McDowell and Chinchilla 2016).

Decolonial theory provides us with several useful tools with which to qualify the nature of power imbalances or inequitable impacts that arise from advanced technologies like AI. One such framework identifies metropoles—the centres of power—and their peripheries that hold relatively less power and contest the metropole's authority, participation and legitimacy in shaping everyday life (Champion 2005). Dependency theory expands on this framework, by tying colonial histories to present day underdevelopment and continued economic imbalance between countries and regions, as well as tying resulting dependencies to historic metropole and periphery dynamics (Champion 2005). Using the lens of metropole and periphery, we can identify contemporary practices in AI development partially as features of colonial continuities from states and governments. Similarly, today's technology corporations could be described as metropoles of technological power with civic society and consumers sitting at the periphery.

Metropole-periphery dichotomies are interpretive models that if not used carefully can reduce the reality of lived experiences to overly simplified binaries of "West and the rest", "North and South" and "powerful and oppressed" (McClintock 1992; Stoler 2008; Thrush 2008), exposing some of the limitations of decolonial theory. In addition, grand historical meta-narratives of injustice, contending with the theoretical idea of "global" and "speaking for the oppressed" (Pappas 2017), are pitfalls that must be avoided. A needed balance can be found by incorporating other modes of decolonial thought, such as contrapuntal analysis (Said 1993), psychodynamic perspectives (Fanon 1961; Nandy 1989), economic analysis (Pollard et al. 2011) and historical and literary criticism (James 1993; Gopal 2019), amongst others. Because of limitations of the theory, we believe it is important to incorporate the wider critical science view introduced in the previous section.

3 Algorithmic Coloniality

By recognising the analogues of territorial and structural coloniality in the digital age, we propose the application of decolonial theory to digital technologies such as AI. Digital spaces—created by the Internet and the increasingly networked systems and devices we use—form digital territories that, like physical spaces, have the propensity to become sites of extraction and exploitation, and thus the sites of digital-territorial coloniality. Digital-structural coloniality also manifests, through to the coloniality of power, the coloniality of power can be observed in digital structures in the form of socio-cultural imaginations, knowledge systems and ways of developing and using technology which are based on systems, institutions, and values which persist from the past and remain unquestioned in the present. As such, emerging technologies like AI are directly subject to coloniality, giving decolonial critical theories a powerful analytical role.

The emerging theories of data colonialism (Thatcher et al. 2016; Ricaurte 2019; Couldry and Mejias 2019a) and data capitalism (Zuboff 2019) recognise this nature of historic continuity and the role of data as the material resource that is exploited for economic expansion. Ricaurte (2019) develops a theoretical model that analyses the coloniality of technological power through data, examining data-centric epistemologies as an expression of the coloniality of power, in how they impose "ways of being, thinking, and feeling that leads to the expulsion of human beings from the social order, denies the existence of alternative worlds and epistemologies, and threatens life on Earth" (Ricaurte 2019). Couldry and Mejias U.A. (2019b) further expand on the colonial continuities of extraction and exploitation of land, labour and relations through digital infrastructure. This larger area of technological coloniality is further developed by scholars in areas of intersectional data feminism (D'Ignazio and Klein 2020), critical race theory (Benjamin 2019), decolonisation of technology (Awori et al. 2016), new data epistemologies (Milan and Van der Velden 2016), and environmental sustainability and justice (Røpke 2001; Ricaurte 2019; Gallopin 1992).

We use the term *algorithmic coloniality* to build upon data colonialism in the context of the interactions of algorithms across societies, which impact the allocation

of resources, human socio-cultural and political behaviour and extant discriminatory systems. We also begin to examine how coloniality features in algorithmic decision-making systems as they generate new labour markets, impact geopolitical power dynamics and influence ethics discourse. In the following section, we introduce the language of decoloniality to the current discourse on fairness, accountability and transparency in algorithmic systems, as well as introduce a taxonomy of decolonial foresight: institutionalised algorithmic oppression, algorithmic exploitation and algorithmic dispossession. Within these forms of decolonial foresight, we present a range of use cases that we identify as sites of coloniality: algorithmic decision systems, ghost work, beta-testing, national policies and international social development. By sites of coloniality, we mean cases that exhibit structural inequalities that can be contextualised historically as colonial continuities. These sites of coloniality help identify where empirical observation departs from the current theoretical frameworks of power in AI, which by-and-large are ahistorical. By using these sites to address the clash of theory and empiricism, we argue that discussions of power and inequality as related to AI cannot be ahistorical and are incomplete if they fail to recognise colonial continuities.

3.1 Algorithmic Oppression

Algorithmic oppression extends the unjust subordination of one social group and the privileging of another—maintained by a "complex network of social restrictions" ranging from social norms, laws, institutional rules, implicit biases and stereotypes (Taylor 2016)—through automated, data-driven and predictive systems. The notion of algorithmic or automated forms of oppression has been studied by scholars such as Noble (2018) and Eubanks (2018). The following examples will make initial connections between instances of algorithmic oppression across geographies, and identify the role of decolonial theory in this discourse.

3.1.1 Site 1: Algorithmic Decision Systems

Predictive systems leveraging AI have led to the formation of new types of policing and surveillance and access to government services, and reshaped conceptions of identity and speech in the digital age. Such systems were developed with the ostensible aim of providing decision-support tools that are evidence-driven, unbiased and consistent. Yet, evidence of how these tools are deployed shows a reality that is often the opposite. Instead, these systems risk entrenching historical injustice and amplify social biases in the data used to develop them (Benjamin 2019).

Evidence of such instances are abundant. As an example, digital human rights concerns have been widely raised: in Singapore's application of facial recognition in CCTV through the Lamppost as a Platform initiative (LaaP) (Johnston 2019), New Delhi's CMAPS predictive policing system (Marda and Narayan 2020), India's Aadhaar identity system (Siddiqui and Singh 2015), the Kenyan Huduma Namba digital/biometric identity system (Nyawa 2019), and the welfare interventions for Māori children by the New Zealand government (Vaithianathan et al. 2013; Gavighan et al. 2019). The impact of predictive algorithms on people, from everyday citizens to the

most vulnerable, highlights the need for diversified and contextualised approaches to issues of justice and fairness in automated systems.

Algorithmic decision systems (Isaac 2017) are increasingly common within the US criminal justice system despite significant evidence of shortcomings, such as the linking of criminal datasets to patterns of discriminatory policing (Angwin et al. 2016; Lum and Isaac 2016; Richardson et al. 2019). Beyond the domain of criminal justice, there are numerous instances of predictive algorithms perpetuating social harms in everyday interactions, including examples of facial recognition systems failing to detect Black faces and perpetuating gender stereotypes (Buolamwini and Gebru 2018; Keyes 2018; Stark 2019), hate speech detection algorithms identifying Black and queer vernacular as "toxic" (Sap et al. 2019), new recruitment tools discriminating against women (Dastin 2018), automated airport screening systems systematically flagging trans bodies for security checks (Costanza-Chock 2018) and predictive algorithms used to purport that queerness can be identified from facial images alone (Agüera y Arcas et al. 2018).

The current discourse on sociotechnical systems[3], under which many of these cases are discussed, can be further enriched if these inequities are historically contextualised in global systems of racial capitalism, class inequality and heteronormative patriarchy, rooted in colonial history (Bhattacharyya and et al. 2018; Sokoloff and Pincus 2008). In the case of racial capitalism, similarly to Ricaurte (2019), Couldry and Mejias (2019a) and Milan and Treré (2019),we propose that institutionalised harms replicated by automated decision-making tools should be understood as continuous to, and inextricably linked with, "histories of racist expropriation", and that "only by tracking the interconnections between changing modes of capitalism and racism that we can hope to address the most urgent challenges of social injustice" (Bhattacharyya and et al. 2018).

A decolonial framework helps connect instances of algorithmic oppression to wider socio-political and cultural contexts, enabling a geographically, historically and intersectionally expansive analysis of risks and opportunities pertaining to AI systems. Notably, it allows for the analysis to move beyond North American or European identity frameworks or definitions of harms. By connecting instances of algorithmic oppression across geographies, new approaches that consider alternative possibilities of using technology in socially complex settings in more critical and considered ways will emerge, and so too will designs that incorporate inclusive and well-adapted mechanisms of oversight and redress from the start.

3.2 Algorithmic Exploitation

Algorithmic exploitation considers the ways in which institutional actors and industries that surround algorithmic tools take advantage of (often already marginalised) people by unfair or unethical means, for the asymmetrical benefit of these industries. The following examples examine colonial continuities in labour practices and scientific experimentation in the context of algorithmic industries.

[3]For example, see ACM Conference on Fairness, Accountability, and Transparency.

3.2.1 Site 2: Ghost Workers

Many of the recent successes in AI are possible only when the large volumes of data needed are annotated by human experts to expose the common sense elements that make the data useful for a chosen task. The people who do this labelling for a living, the so-called "ghost workers" (Gray and Suri 2019), do this work in remote settings, distributed across the world using online annotation platforms or within dedicated annotation companies. In extreme cases, the labelling is done by prisoners (Hao 2019) and the economically vulnerable (Yuan 2018), in geographies with limited labour laws. This is a complicated scenario. On one hand such distributed work enables economic development, flexibility in working and new forms of rehabilitation. On the other, it establishes a form of knowledge and labour extraction, paid at very low rates, and with little consideration for working conditions, support systems and safeties.

A decolonial lens shifts our view towards understanding how colonial history affects present-day labour regulation and enforcement (Ronconi 2015), and how the capacity to mobilise production and outsource services across borders allows industries to take advantage of present-day post-colonial economic inequalities in order to "reorganize production in ways and places that reduce manufacturing costs and enhance corporate profit" (Gomberg-Muñoz 2018; Wallerstein 1987). Logics of colonial extraction and exploitation have "mutated but also maintain continuity in the present day", supporting post-colonial economic inequalities which have been empirically demonstrated to be tied to historic colonial activity (Bruhn and Gallego 2012; Fanon 1961). Data generation and processing presents opportunities for extraction and excavation within data mining industries, which are arguably "ingrained in practices and techniques of extraction [and] is a kind of colonial imprint" (Mezzadra and Neilson 2017), as demonstrated in part by the location of many ghost workers in previously colonised geographies.

3.2.2 Site 3: Beta-testing

There is a long and well-documented history on the exploitation of marginalised populations for the purpose of scientific and technological progress. Colonies of the British empire "provided a laboratory for experimenting with new forms of medical and scientific practice" (Senior 2018; Tilley 2014). There has been a historic continuity of scientific experimentation on African Americans, from early experimentation on black enslaved women and infants in the nineteenth century that is foundational to the field of gynaecology (Washington 2006), to the Tuskegee syphilis study (Brandt 1978). Such experimental practices continue to colour the establishment of socio-economic development schemes in previously colonised countries, often by former colonisers (Bonneuil 2000).

It is with this historic lens that we examine the practice of beta-testing, which is the testing and fine-tuning of early versions of software systems to help identify issues in their usage in settings with real users and use cases. In the testing of predictive systems, we find several clearly exploitative situations, where organisations use countries outside of their own as testing grounds—specifically because they lack

pre-existing safeguards and regulations around data and its use, or because the mode of testing would violate laws in their home countries (UNCTAD 2013). This phenomenon is known as *ethics dumping*: the export of harms and unethical research practices by companies to marginalised and vulnerable populations or to low- and middle-income countries, and which often aligns "with the old fault lines of colonialism" (Schroeder et al. 2018). The counterpoint to ethics dumping is *ethics shirking*: what is not done to protect people when harms emerge beyond what is demanded from legal or regulatory frameworks (Floridi 2019).

As an example, Cambridge Analytica (CA) elected to beta-test and develop algorithmic tools for the 2017 Kenyan and 2015 Nigerian elections, with the intention to later deploy these tools in US and UK elections. Kenya and Nigeria were chosen in part due to the weaker data protection laws compared to CA's base of operations in the United Kingdom—a clear example of ethics dumping. These systems were later found to have actively interfered in electoral processes and worked against social cohesion (Nyabola 2018). A critical decolonial approach would, for example, lead us to ask early on why the transgression of democratic processes by companies such as CA only gained international attention and mobilisation after beginning to affect Western democratic nations.

In another case of beta-testing, the deployment of predictive algorithms for child welfare interventions by the New Zealand government initially targeted Māori, the indigenous people of New Zealand, who have long experienced institutional racism (Vaithianathan et al. 2013; Gavighan et al. 2019). Analogously, the data analytics firm Palantir was found to have experimentally deployed predictive algorithms in the city of New Orleans (in concert with the police department) without public approval. These tools were used to target specific individuals and neighbourhoods for police surveillance, and disproportionately impacted African-Americans (Bullington and Lane 2018). These are all cases that cannot be viewed ahistorically, e.g. for African Americans there is a historic continuity from the nineteenth century gynaecology experimentation, to the twentieth century Tuskegee experiments, to the twenty-first century predictive policing and beta-testing.

The perspective of historic continuity provided by decolonial theory raises important questions of accountability, responsibility, contestation and recourse, which become increasingly necessary in entangled settings of low regulation, combined with deficits of localised expertise and contextualised historic knowledge within firms expanding into new markets. Risks are likely to arise if we neglect to explore the current variation of ethical standards based on identity and geography, as well as how algorithms and automated systems interact with existing social stratification at both local and global levels.

3.3 Algorithmic Dispossession

Algorithmic dispossession, drawing from Harvey (2004) and Thatcher et al. (2016), describes how, in the growing digital economy, certain regulatory policies result in a centralisation of power, assets, or rights in the hands of a minority and the deprivation of power, assets or rights from a disempowered majority. The following examples

examine this process in the context of international AI governance (policy and ethics) standards, and AI for international social development.

3.3.1 Site 4: National Policies and AI Governance

Power imbalances within the global AI governance discourse encompass issues of data inequality and data infrastructure sovereignty, but also extend beyond this. We must contend with questions of *who* any AI regulatory norms and standards are protecting, who is empowered to project these norms and the risks posed by a minority continuing to benefit from the centralisation of power and capital through mechanisms of dispossession (Thatcher et al. 2016; Harvey 2004). As Jasanoff and Hurlbut (2018) remind us, we must be mindful of "who sits at the table, what questions and concerns are sidelined and what power asymmetries are shaping the terms of debate".

A review of the global landscape of AI ethics guidelines (Jobin et al. 2019) pointed out the "under-representation of geographic areas such as Africa, South and Central America and Central Asia" in the AI ethics debate. The review observes a power imbalance wherein "more economically developed countries are shaping this debate more than others, which raises concerns about neglecting local knowledge, cultural pluralism and the demands of global fairness". A similar dynamic is found when we examine the proliferation of national policies on AI in countries across the world (Dutton 2018). In some views, this is a manifestation of a new type of geopolitics amongst "AI superpowers" (Lee 2018), and a rise of "AI nationalism", where nations wrangle to spread a preferred view of policy, applied approaches and technical services (Hogarth 2018; Edgerton 2007b). We are quickly led to one possible scene of coloniality by Lee (2017): "Unless they [developing countries] wish to plunge their people into poverty, they will be forced to negotiate with whichever country supplies most of their AI software—China or the United States—to essentially become that country's economic dependent". It can be argued that the agency of developing countries is in these ways undermined, where they "cannot act unilaterally to forge their own rules"and cannot expect prompt protection of their interests (Pathways for Prosperity 2019).

Such concerns were demonstrated at the 2019 G20 summit, where a number of developing countries including India, Indonesia and South Africa refused to sign the Osaka Track, an international declaration on data flows (Kanth 2019), because the interests, concerns and priorities of these countries were not seen to be represented in the document. The undermining of interests and agency of developing countries is also a relevant issue vis à vis the OECD AI Principles (OECD 2019). As these guidelines are adopted and enforced by partner countries around the world, we see analogous concerns surfacing around exclusionary path dependencies and first-mover advantages (Pathways for Prosperity 2019). Additionally, AI governance guidelines risk being replicated across jurisdictions in a way that may be incompatible with the needs, goals and constraints of developing countries, despite best efforts (Pathways for Prosperity 2019).

There are clear hierarchies of power within these cases of policy development, which can be analysed using the aforementioned metropole-periphery model. It is

metropoles (be it government or industry) who are empowered to impose normative values and standards, and may do so at the "risk of forestalling alternative visions" (Greene et al. 2019). A metropole-periphery model draws attention to the need to represent values, interests, concerns and priorities of resource-constrained countries in AI governance processes, as well as the historic dynamics that prevent this. Decolonial theory offers AI policy makers a framework to interrogate imbalances of power in AI policy discourse, understand structural dependencies of developing countries, question ownership of critical data infrastructures and assess power imbalances in product design/development/deployment of computational technologies (Irani et al. 2010) as well as the unequal distribution of risks and economic benefits.

3.3.2 Site 5: International Social Development

Much of the current policy discourse surrounding AI in developing countries is in economic and social development where advanced technologies are propounded as solutions for complex developmental scenarios, represented by the growing areas of AI for Good and AI for the Sustainable Development Goals (AI4SDGs) (Vinuesa et al. 2020; Floridi et al. 2018; Tomašev et al. 2020). In this discourse, Green (2019) proposes that "good isn't good enough", and that there is a need to expand the currently limited and vague definitions within the computer sciences of what "social good" means.

To do so, we can draw from existing analysis of ICT for Development, which are often based on historical analysis and decolonial critique (Irani et al. 2010; Toyama 2015). These critiques highlight concerns of dependency, dispossession or ethics dumping and shirking, as discussed earlier (Schroeder et al. 2018). Such critiques take renewed form as AI is put forward as a needed tool for social development. Where a root cause of failure of developmental projects lies in default attitudes of paternalism, technological solutionism and predatory inclusion, e.g. "surveillance humanitarianism" (Latonero 2019; Vinuesa et al. 2020), decolonial thinking shifts our view towards systems that instead promote active and engaged political community. This implies a shift towards the design and deployment of AI systems that is driven by the the agency, self-confidence and self-ownership of the communities they work for, e.g. adopting co-development strategies for algorithmic interventions alongside the communities they are deployed in (Katell et al. 2020).

Co-development is one potential strategy within a varied toolkit supporting the socio-political, economic, linguistic and cultural relevance of AI systems to different communities, as well as shifting power asymmetries. A decolonial view offers us tools with which to engage a reflexive evaluation and continuous examination of issues of cultural encounter, and a drive to question the philosophical basis of development (Kiros 1992). With a self-reflexive practice, initiatives that seek to use AI technologies for social impact can develop the appropriate safeguards and regulations that avoid further entrenching exploitation and harm, and can conceptualise long-term impacts of algorithmic interventions with historical continuities in mind.

4 Tactics for a Decolonial AI

By fusing the fields of artificial intelligence and decolonial theories, we can take advantage of historical hindsight to develop new tools of foresight and practice. In so doing, we can establish a decolonial AI that can re-create the field of artificial intelligence in ways that strengthens its empirical basis, while anticipating and averting algorithmic colonialism and harm.

The five sites of coloniality in the previous section cast the applications of AI research (its products and predictions—AI as object) and the structures that support it (data, networks and policies—AI as subject) as expressions of the coloniality of power (Quijano 2000, Mignolo 2007; Maldonado-Torres 2007; Ndlovu-Gatsheni 2015), and of technological power (Ricaurte 2019; Couldry and Mejias 2019a; Ali 2016). This leads us to seek the decolonisation of power, whose aim is dismantle harmful power asymmetries and concepts of knowledge, turning us instead towards a "pluriversal epistemology of the future" (Mignolo 2012) that unlike universalims, acknowledges and supports a wider radius of socio-political, ecological, cultural and economic needs.

In this final section, we aim to develop sets of *tactics* for the future development of AI, which we believe open many areas for further research and action. Tactics do not lead to a conclusive solution or method, but instead to the "contingent and collaborative construction of other narratives" (Philip et al. 2012). Our tactics resonate with the proposals for reforming epistemic practice articulated by many other scholars. Couldry and Mejias (2019a) put forward a vision for decolonising data relations by exploring six tasks—reframing what data is for, restoring well-being, naming alternative world views, gendering, protecting and creating new forms of social relations—that must all be oriented towards social goals. Ricaurte (2019) points to needed change in data governance and regimes, addressing technological sovereignty and agency, and addressing the impact of technological systems on ecological systems and the need to imagine alternative digital futures. Benjamin (2019) asks us to retool solidarity and reimagine justice, by rethinking design, developing coded equity audits and developing abolitionist tools that reimagine technology.

We submit three tactics for future AI design—supporting a critical technical practice of AI, establishing reciprocal engagements and reverse pedagogies and the renewal of affective and political community—based on lessons of resistance and recovery from historical and decolonial criticism, and grounded within already existing work that shows how these tactics might be enacted in practice.

4.1 Towards a Critical Technical Practice of AI

The basis of decolonial AI rests in a self-reflexive approach to developing and deploying AI that recognises power imbalances and its implicit value systems. It is exactly this type of framework that was developed by Agre (1997), who described a shift towards a *Critical Technical Practice of AI* (CTP). Critical technical practices take a middle ground between the technical work of developing new AI algorithms and the reflexive work of criticism that uncovers hidden assumptions and alternative ways of working. CTP has been widely influential, having found an important place

in human-computer interactions (HCI) and design (Dourish et al. 2004; Sengers et al. 2006). By infusing CTP with decoloniality, we can place a productive pressure on our technical work, moving beyond good-conscience design and impact assessments that are undertaken as secondary tasks, to a way of working that continuously generates provocative questions and assessments of the politically situated nature of AI.

The role of *practice* in this view is broad by necessity. Recent research, in both AI and Science and Technology Studies (STS), highlights the limitations of purely technological approaches to addressing the ethical and social externalities of AI. Yet, technical approaches can meaningfully contribute when they appropriately reflect the values and needs of relevant stakeholders and impacted groups (Selbst et al. 2019). This context-aware technical development that CTP speaks to—which seeks to consider the interplay between social, cultural and technical elements—is often referred to as heterogeneous engineering (Law and et al. 1987). As a result, a heterogeneous-critical practice must encompass multiple approaches for action: in research, organising, testing, policy and activism. We explore five topics constituting such a practice: algorithmic fairness, AI safety, equity and diversity, policy-making, and AI as a decolonising tool.

Fairness Research in *algorithmic fairness* (Nissenbaum 2001; Dwork et al. 2012; Barocas and Selbst 2016) has recognised that efforts to generate a fair classifier can still lead to discriminatory or unethical outcomes for marginalised groups, depending on the underlying dynamics of power; because a "true" definition of fairness is often a function of political and social factors. Quijano (2000) again speaks to us, posing questions of who is protected by mainstream notions of fairness, and to understand the exclusion of certain groups as "continuities and legacies of colonialism embedded in modern structures of power, control, and hegemony". Such questions speak to a critical practice whose recent efforts, in response, have proposed fairness metrics that attempt to use causality (Chiappa and Isaac 2019; Mitchell et al. 2018; Nabi and Shpitser 2018; Madras et al. 2019) or interactivity (Canetti et al. 2019; Jung et al. 2019) to integrate more contextual awareness of human conceptions of fairness.

Safety The area of *technical AI safety* (Amodei et al. 2016; Raji and Dobbe 2020) is concerned with the design of AI systems that are safe and appropriately align with human values. The philosophical question of value alignment arises, identifying the ways in which the implicit values learnt by AI systems can instead be aligned with those of their human users. A specification problem emerges when there is a mismatch between the ideal specification (what we want an AI system to do) and the revealed specification (what the AI system actually does). This again raises questions that were posed in the opening of whose values and goals are represented, and who is empowered to articulate and embed these values—introducing discussions of utilitarian, Kantian and volitional views on behaviour, and on the prevention and avoidance of undesirable and unintended consequences (Gabriel 2020). Of importance here, is the need to integrate discussions of social safety alongside questions of technical safety.

Diversity With a critical lens, efforts towards greater equity, diversity and inclusion (EDI) in the fields of science and technology are transformed from the prevailing

discourse that focuses on the business case of building more effective teams or as being a moral imperative (Rock and Grant 2016), into *diversity as a critical practice* through which issues of homogenisation, power, values and cultural colonialism are directly confronted. Such diversity changes the way teams and organisations think at a fundamental level, allowing for more intersectional approaches to problem-solving to be taken (D'Ignazio and Klein 2020).

Policy There is growing traction in *AI governance* in developing countries to encourage localised AI development, such as the initiatives by UNESCO, UN Global Pulse's AI policy support in Uganda and Ghana (ITU 2019) and Sierra Leone's National Innovation & Digital Strategy (DSTI 2019), or in structuring protective mechanisms against exploitative or extractive data practices (Gray and Suri 2019). Although there are clear benefits to such initiatives, international organisations supporting these efforts are still positioned within metropoles, maintaining the need for self-reflexive practices and considerations of wider political economy (Pathways for Prosperity 2019).

Resistance The *technologies of resistance* have often emerged as a consequence of opposition to coloniality, built by self-organising communities to "bypass dynamics and control of the state and corporations" (Steiner 1994; Milan 2013). A renewed critical practice can also ask the question of whether AI can itself be used as a decolonising tool, e.g. by exposing systematic biases and sites of redress. For example, Chen et al. (2019) instantiate this idea of using AI to assess systemic biases in order to reduce disparities in medical care, by studying mortality and 30-day psychiatric readmission with respect to race, gender, and insurance payer type as a proxy for socioeconomic status. Furthermore, although AI systems are confined to a specific sociotechnical framing, we believe that they can be used as a decolonising tool while avoiding a techno-solutionism trap. When AI systems can be adapted to locally specific situations in original ways, they can take a renewed role as "creole technologies" that find positive and distinctive use at scale, and outside their initially conceived usage (Edgerton 2007a).

4.2 Reciprocal Engagements and Reverse Tutelage

Research in post-colonial studies increasingly highlights the essential role that colonised peoples themselves, through insurgence, activism and organisation, had in changing the colonial view in the metropole (Gopal 2019; Gandhi 2006). Despite colonial power, the historical record shows that colonialism was never only an act of imposition. In a reversal of roles, the metropole often took lessons from the periphery, establishing a reverse tutelage between centre and periphery. A modern critical practice would seek to use this decolonial imperative to develop a double vision: actively identifying centres and peripheries that make reverse tutelage and the resulting pedagogies of reciprocal exchange part of its foundations, while also seeking to undo colonial binarisms.

Reverse tutelage directly speaks to the philosophical questions of what constitutes knowledge. There remains a tension between a view of knowledge as absolute and

of data that, once enough is collected, allows us to form complete and encompassing abstractions of the world, versus a view of knowledge that is always incomplete and subject to selections and interpretation under differing value systems. These oppositional views of knowledge have been explored in different ways, such as in the important anthropological work by Forsythe (1993, 2001) on knowledge in AI, in the genealogy of statistics as "the moral science" (Hacking 2015) and through "new data epistemologies" (Milan and Van der Velden 2016). Deciding what counts as valid knowledge, what is included within a dataset and what is ignored and unquestioned is a form of power held by AI researchers that cannot be left unacknowledged. It is in confronting this condition that decolonial science, and particularly the tactic of reverse tutelage, makes its mark. We put forward three modes—of dialogue, documentation and design—through which reciprocal tutelage can be enacted.

Dialogue Reverse pedagogies create a decolonial shift from paternalistic towards solidaristic modes of working that can be achieved by systems of meaningful *intercultural dialogue*. Such dialogue is core to the field of intercultural digital ethics, which asks questions of how technology can support society and culture, rather than becoming an instrument of cultural oppression and colonialism (Capurro 2018). Intercultural ethics emphasises the limitations and coloniality of universal ethics—dominant rather than inclusive ethical frameworks—and finds an alternative in pluralism, pluriversal ethics and local designs (Escobar 2011; Ess 2006). One approach to reverse pedagogies is invoked by Arora (2019) in the field of privacy research, by interrogating the empirical basis of privacy studies, and calling for an "epistemic disobedience" and a reliance on shifting roles of the metropole and periphery.

Documentation New frameworks have been developed that make explicit the representations of knowledge assumed within a dataset and within deployed AI systems. Data sheets for datasets aim to summarise what is and is not contained within a dataset (Gebru et al. 2018), and similar explicit assessments for AI systems exist using the model cards framework (Mitchell et al. 2019). The example in Mitchell et al. (2019) on toxicity scoring provides a simple and powerful example of reverse pedagogy, wherein affected users exposed the system's limitations that led to documented improvements in its subsequent releases.

Design There is now also a growing understanding of approaches for meaningful community-engaged research (Mikesell et al. 2013), using frameworks like the IEEE Ethically Aligned Design (IEEE Global Initiative 2016), technology policy design frameworks like Diverse Voices (Young et al. 2019) and mechanisms for the co-development of algorithmic accountability through participatory action research (Katell et al. 2020). The framework of citizens' juries have also been used to gain insight into the general public's understanding of the role and impact of AI (Balaram et al. 2018).

A critical viewpoint may not have been the driver of these solutions, and these proposals are themselves subject to limitations and critique, but through an ongoing process of criticism and research, they can lead to powerful mechanisms for reverse tutelage in AI design and deployment.

4.3 Renewed Affective and Political Communities

How we build a critical practice of AI depends on the strength of political communities to shape the ways they will use AI, their inclusion and ownership of advanced technologies, and the mechanisms in place to contest, redress and reverse technological interventions. The systems we described in Section 3, although ostensibly developed to support human decision-makers and communities, failed to meaningfully engage with the people who would be the targets of those systems, cutting off these avenues of ownership, inclusion and justice. The historical record again shows that these situations manifest through paternalistic thinking and imbalances in authority and choice, produced by the hierarchical orders of division and binarisation established by coloniality (Gopal 2019; Said 1993; Fanon 1967; Nandy 1989). The decolonial imperative asks for a move from attitudes of technological benevolence and paternalism towards solidarity. This principle enters amongst the core of decolonial tactics and foresight, speaking to the larger goal of decolonising power.

The challenge to solidarity lies in how new types of political community can be created that are able to reform systems of hierarchy, knowledge, technology and culture at play in modern life. One tactic lies in embedding the tools of decolonial thought within AI design and research. Contrapuntal analysis (Said 1993) is one important critical tool that actively leads us to expose the habits and codifications that embed questionable binarisms—of metropole and periphery, of West and the rest, of scientists and humanist, of natural and artificial—in our research and products. Another tactic available to us lies in our support of grassroots organisations and in their ability to create new forms of affective community, elevate intercultural dialogue and demonstrate the forms of solidarity and alternative community that are already possible. Many such groups already exist, particularly in the field of AI, such as Data for Black Lives (Goyanes 2018), the Deep Learning Indaba (Gershgorn 2019), Black in AI and Queer in AI, and are active across the world.

The advantage of historical hindsight means that we can now recover the principles of living that were previously made incompatible with life by colonial binaries. Friendship quickly emerges as a "lost trope of anticolonial thought" (Gandhi 2006). This is a political friendship that has been expanded in many forms: by El Khayat and Khatibi (2010) using the concept of *aimance*, in the politics of friendship (Derrida 1993) and as *affective communities* (Gandhi 2006) in which developers and users seek alliances and connection outside possessive forms of belonging. Other resources are also available: within the principles of political love (Fanon 1967; Zembylas 2017; Butorac 2018), in moral philosophies that recognise both material and immaterial development (Kiros 1992) and in philosophies such as Ubuntu (Ramose 1999) and Oneness (Wong 2012). And the decolonial principles we described—moves from benevolence to solidarity, enabling systems of reverse tutelage, harnessing technological solutions, creating a critical technical practice—when combined, shape the creation of these new types of political and affective community. These ideas are relevant to the work of AI at its most fundamental because recalling the use cases in Section 3, AI is shaped by, and shapes the evolution of contemporary political community.

Finally, these views of AI taken together lead us quickly towards fundamental philosophical questions of what it is to be human—how we relate and live with each other in spaces that are both physical and digital, how we navigate difference and transcultural ethics, how we reposition the roles of culture and power at work in daily life—and how the answers to these questions are reflected in the AI systems we build. Here alone is there an ongoing need for research and action, and to which historical hindsight and technological foresight can make significant contributions.

5 Conclusions

This paper aims to provide a perspective on the importance of a critical science approach, and in particular of decolonial thinking, in understanding and shaping the ongoing advances in AI. Despite these fields having operated mostly apart, our hope is that decolonial theories will expand the practitioner's ability, including ourselves, to ask critical questions of technology research and design, to imagine alternative realities, to question who is developing AI and where, and to examine the roles of culture and power embedded in AI systems; reinvigorating Agre's (1997) vision of a critical technical practice of AI.

Operationalising this critical practice will require not only foresight and case study research, but also approaches that support new research cultures, along with innovative technical research in domains such as fairness, value alignment, privacy and interpretability. Moreover, there is a strong need to develop new methodologies for inclusive dialogue between stakeholders in AI development, particularly those in which marginalised groups have meaningful avenues to influence the decision-making process, avoiding the potential for predatory inclusion and continued algorithmic oppression, exploitation and dispossession.

Any commitment to building the responsible and beneficial AI of the future ties us to the hierarchies, philosophy and technology inherited from the past, and a renewed responsibility to the technology of the present. To critically engage with that inheritance, to avert algorithmic coloniality, to reveal the relations of power that underlie technology deployment, to recover the principles of living that were previously made incompatible with life and to create new forms of political and affective community, are the tasks for which decolonial theories will provide key resources—whether these be through heterogeneous practice, intercultural dialogue, creole AIs, reverse tutelage or a politics of friendship and love. The years ahead will usher in a wave of new scientific breakthroughs and technologies driven by AI, making it incumbent upon AI communities to strengthen the social contract through ethical foresight and the multiplicity of intellectual perspectives available to us, aligned with the goal of promoting beneficence and justice for all.

Acknowledgements We thank our reviewers and editor whose positive feedback and mentorship has improved this paper. We are grateful to Desha Osborne, Irene Solaiman, Momin Malik and Vafa Ghazavi, who have influenced our thinking. We are also grateful to many of our colleagues for their support, including Ben Coppin, Brittany Smith, Courtney Biles, Dorothy Chou, Heiko Strathmann, Iason Gabriel, James Besley, Kelly Clancy, Koray Kavukcuoglu, Laura Weidinger, Murray Shanahan and Sean Legassick.

Disclaimer Any opinions presented in this paper represent the personal views of the authors and do not necessarily reflect the official policies or positions of their organisations.

References

Agre, P. (1997). Toward a critical technical practice: lessons learned in trying to reform AI. In Bowker, G., Star, S., Gasser, L., Turner, W. (Eds.) *Social science, technical systems and cooperative work: beyond the great divide, psychology press.* pp. 131–157.

Agüera y Arcas, B., Todorov, A., Mitchell, M. (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? Medium https://link.medium.com/GO7FJgFgM1.

Ali, S.M. (2016). A brief introduction to decolonial computing. *XRDS: Crossroads The ACM Magazine for Students*, *22*(4), 16–21.

AMA (2018). Augmented intelligence in health care H-480.940. American Medical Association PolicyFinder.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2016). Concrete problems in AI safety. arXiv:160606565.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016a). Machine bias. Propublica, May 23:2016.

Angwin, J., Larson, J., Mattu, S., Kirchner, L. (2016). Machine bias. There's software used across the country to predict future criminals. and it's biased against blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arora, P. (2019). Decolonizing privacy studies. *Television & New Media*, *20*(4), 366–378.

Ashcroft, B. (2006). The post-colonial studies reader. Taylor & Francis, Tiffin, H.

Asilomar Meeting (2017). Asilomar AI principles. https://futureoflife.org/ai-principles/.

Awori, K., Bidwell, NJ., Hussan, TS., Gill, S., Lindtner, S. (2016). Decolonising technology design. In *Proceedings of the first African conference on human computer interaction,* pp. 226–228.

Balaram, B., Greenham, T., Leonard, J. (2018). Artificial intelligence: real public engagement. London: RSA. https://wwwthersaorg/discover/publications-and-articles/reports/artificial-intelligence-realpublic-engagement.

Barocas, S., & Selbst, A.D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Belmont Commission (1978). The Belmont report: ethical principles and guidelines for the protection of human subjects of research, vol 1. United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Benjamin, R. (2019). *Race after technology: abolitionist tools for the new Jim Code*. New York: John Wiley & Sons.

Bhambra, G.K., Nisancioglu, K., Gebrial, D. (2018). *Decolonizing the university*. London: Pluto Press.

Bhattacharyya, G., et al. (2018). *Rethinking racial capitalism: questions of reproduction and survival*. Maryland: Rowman & Littlefield International.

Boden, M.A. (2018). *Artificial intelligence: a very short introduction*. London: Oxford University Press.

Bonneuil, C. (2000). Development as experiment: science and state building in late colonial and postcolonial Africa, 1930-1970. *Osiris*, *15*, 258–281.

Brandt, A.M. (1978). *Racism and research: the case of the Tuskegee syphilis study*, (pp. 21–29). New York: Hastings Center Report.

Bruhn, M., & Gallego, F.A. (2012). Good, bad, and ugly colonial activities: do they matter for economic development?. *Review of Economics and Statistics*, *94*(2), 433–461.

Bueter, A. (2015). The irreducibility of value-freedom to theory assessment. *Studies in History and Philosophy of Science Part A*, *49*, 18–26.
Bullington, J., & Lane, E. (2018). How a tech firm brought data and worry to New Orleans crime fighting. The New Orleans Times-Picayune. https://www.nola.com/news/crime_police/article_33b8bf05-722f-5163-9a0c-774aa69b6645.html.
Buolamwini, J., & Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency,* pp 77–91.
Butorac, S.K. (2018). Hannah Arendt, James Baldwin, and the politics of love. *Political Research Quarterly*, *71*(3), 710–721.
Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., Smith, A. (2019). From soft classifiers to hard decisions: how fair can we be? In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 309–318). New York: ACM.
Capurro, R. (2018). Intercultural information ethics. In *Localizing the Internet* (pp. 19–38). Munich: Wilhelm Fink Verlag.
Champion, T. (2005). Metropole and margin: the dependency theory and the political economy of the Solomon Islands, 1880–1980. In *Centre and periphery* (pp. 43–60). Evanston: Routledge.
Chen, I.Y., Szolovits, P., Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care?. *AMA journal of ethics*, *21*(2), 167–179.
Chiappa, S., & Isaac, W.S. (2019). A Causal Bayesian Networks Viewpoint on Fairness. In Kosta, E., Pierson, J., Slamanig, D., Fischer-Hübner, S., Krenn, S. (Eds.) *Privacy and identity management. Fairness, accountability, and transparency in the age of big data. Privacy and identity 2018. IFIP advances in information and communication technology*, (Vol. 547 pp. 3–20). Cham.
Coates, J.F., & at al. (1985). Foresight in federal government policy making. *Futures Research Quarterly*, *1*(2), 29–53.
Costanza-Chock, S. (2018). Design justice, AI, and escape from the matrix of domination. Journal of Design and Science.
Couldry, N., & Mejias, U.A. (2019a). *The costs of connection: how data is colonizing human life and appropriating it for capitalism*. Stanford: Stanford University Press.
Couldry, N., & Mejias U.A. (2019b). Data colonialism: rethinking big data's relation to the contemporary subject. *Television & New Media*, *20*(4), 336–349.
Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. San Francisco, CA: Reuters Retrieved on October 9:2018.
Derrida, J. (1993). Politics of friendship. *American Imago*, *50*(3), 353–391.
D'Ignazio, C., & Klein, L.F. (2020). *Data feminism*. Cambridge: MIT Press.
DiSalvo, C. (2012). *Adversarial design (design thinking, design theory)*. Cambridge: MIT Press.
Douglas, H. (2007). Rejecting the ideal of value-free science. In Kincaid, H., Dupré, J., Wylie, A. (Eds.) *Value-free science: ideals and illusions?* chap 6 (pp. 120–141). Oxford: Oxford university press.
Dourish, P., & Mainwaring, S.D. (2012). Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing* (pp. 133–142). New York: ACM.
Dourish, P., Finlay, J., Sengers, P., Wright, P. (2004). Reflective HCI: towards a critical technical practice. In *Conference on human factors in computing systems: CHI'04 extended abstracts on human factors in computing systems*, (Vol. 29 pp. 1727–1728).
Directorate of Science, Technology, and Innovation in the Office of the President, Sierra Leone (2019). Sierra Leone National Innovation & Digital Strategy 2019 - 2029. Digitization for all: Identity, Economy, and Governance.
Dutton, T. (2018). An overview of national AI strategies. Medium. https://link.medium.com/jqPZBjs7j2.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel R (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference,* pp 214–226.
Edgerton, D. (2007a). Creole technologies and global histories: rethinking how things travel in space and time. *Journal of History of Science and Technology*, *1*(1), 75–112.
Edgerton, D.E. (2007b). The contradictions of techno-nationalism and techno-globalism: a historical perspective. New Global Studies 1(1).
El Khayat, R., & Khatibi, A. (2010). Open correspondence: an epistolary dialogue. UNO press translated by Babana-Hampton S, Orlando VK, Vogl M.
Elliott, K.C., & McKaughan, D.J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science*, *81*(1), 1–21.
Escobar, A. (2011). Sustainability: design for the pluriverse. *Development*, *54*(2), 137–140.

Ess, C. (2006). Ethical pluralism and global information ethics. *Ethics and Information Technology*, *8*(4), 215–226.
Eubanks, V. (2018). *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.
Fanon, F. (1961). *The wretched of the earth*. United States: Grove/Atlantic, Inc.
Fanon, F. (1967). *Black skin, white masks*. New York: Grove press.
Feenberg, A. (2017). Critical theory of technology and STS. *Thesis Eleven*, *138*(1), 3–12.
Floridi, L. (2019). Translating principles into practices of digital ethics: five risks of being unethical. *Philosophy & Technology*, *32*(2), 185–193.
Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.
Forsythe, D. (2001). *Studying those who study us: an anthropologist in the world of artificial intelligence*. Stanford: Stanford University Press.
Forsythe, D.E. (1993). Engineering knowledge: the construction of knowledge in artificial intelligence. *Social studies of science*, *23*(3), 445–477.
Freudenthal, E. (2019). Ebola's lost blood: row over samples flown out of africa as' big pharma'set to cash in. The Telegraph.
Friedman, B., Kahn, P.H., Borning, A., Huldtgren, A. (2013). Value sensitive design and information systems. In *Early engagement and new technologies: opening up the laboratory* (pp. 55–95). Berlin: Springer.
Gabriel, I. (2020). Artificial intelligence, values and alignment. arXiv:200109768.
Gallopin, G. (1992). Science, technology and the ecological future of Latin America. *World Development*, *20*(10), 1391–1400.
Gandhi, L. (2006). *Affective communities: anticolonial thought, fin-de-siècle radicalism, and the politics of friendship*. Durham: Duke University Press.
Garcia, E. (2019). The militarization of artificial intelligence: a wake-up call for the global south. Available at SSRN 3452323.
Gavighan, C., Knott, A., Maclaurin, J., Zerilli, J., Liddicoat, J. (2019). *Government use of artificial intelligence in New Zealand*. New Zealand: Law Society.
Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K. (2018). Datasheets for datasets. arXiv:180309010.
Gerrish, S. (2018). *How smart machines think*. Cambridge: MIT Press.
Gershgorn, D. (2019). Africa is building an AI industry that doesn't look like silicon valley. Medium OneZero. https://bit.ly/2SBnQFm.
Gomberg-Muñoz, R.M. (2018). Review essay: law and migrant labor in the 20th century: Ghost workers and global capitalism. PoLAR: Political and Legal Anthropology Review.
Gopal, P. (2019). *Insurgent empire: anticolonial resistance and british dissent*. London: Verso Books.
Goyanes, R. (2018). Data for black lives is an organization using analytics as a tool for social change. Garage Magazine. https://garage.vice.com/en_us/article/kzn4jn/data-for-black-lives-is-an-organization-using-analytics-as-a-tool-for-social-change.
Gray, M.L., & Suri, S. (2019). Ghost work: how to stop silicon valley from building a new global underclass. Eamon Dolan Books.
Green, B. (2019). "Good" isn't good enough. In *NeurIPS workshop on AI for social good*.
Greene, D., Hoffmann, A.L., Stark, L. (2019). Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd hawaii international conference on system sciences*.
Hacking, I. (2015). Biopower and the avalanche of printed numbers. Biopower: Foucault and beyond, 65–80.
Hanna, A., Denton, E., Smart, A., Smith-Loud, J. (2019). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*.
Hao, K. (2019). *An AI startup has found a new source of cheap labor for training algorithms: prisoners*. Cambridge: MIT Tech Review. https://www.technologyreview.com/f/613246/an-ai-startup-has-found-a-new-source-of-cheap-labor-for-training-algorithms/.
Harding, S. (2011). *The postcolonial science and technology studies reader*. Durham: Duke University Press.
Harvey, D. (2004). The 'new' imperialism: accumulation by dispossession. *Socialist Register*, *40*, 63–87.

Hogarth, I. (2018). AI nationalisms. https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism.
IEEE Global Initiative (2016). Ethically aligned design. IEEE Standards v1.
Irani, L., Vertesi, J., Dourish, P., Philip, K., Grinter, R.E. (2010). Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1311–1320). New York: ACM.
Isaac, W.S. (2017). Hope, hype, and fear: the promise and potential pitfalls of artificial intelligence in criminal justice. *Ohio State Journal of Criminal Law*, *15*, 543.
ITU. (2019). *United nations activities on artificial intelligence (AI)*. Geneva: International Telecommunication Union. https://www.itu.int/dms_pub/itu-s/opb/gen/S-GEN-UNACT-2019-1-PDF-E.pdf.
James, C.L.R. (1993). *Beyond a boundary*. Durham: Duke University Press.
Jansen, J. (2019). Decolonisation in universities: the politics of knowledge. Wits University Press.
Jasanoff, S., & Hurlbut, J.B. (2018). A global observatory for gene editing. *Nature*, *555*(7697), 435–437.
Jobin, A., Ienca, M., Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.
Johnston, K. (2019). A comparison of two smart cities: Singapore and Atlanta. *Journal of Comparative Urban Law and Policy*, *3*, 191.
Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., Wu, Z.S. (2019). Eliciting and enforcing subjective individual fairness. arXiv:190510660.
Kanth, D.R. (2019). India boycotts 'Osaka Track' at G20 summit. Live Mint. https://www.livemint.com/news/world/india-boycotts-osaka-track-at-g20-summit-1561897592466.html.
Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Binz, C., Raz, D., Krafft, P. (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 conference on fairness, accountability, and transparency,* pp 45–55.
Keyes, O. (2018). The misgendering machines: trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 88.
Kiros, T. (1992). Moral philosophy and development: the human condition in Africa, vol 61. Ohio Univ Ctr for International Studies.
Latonero, M. (2019). Stop surveillance humanitarianism. New York Times. https://www.nytimes.com/2019/07/11/opinion/data-humanitarian-aid.html.
Laudan, L. (1968). Theories of scientific method from Plato to mach: a bibliographical review. *History of science*, *7*(1), 1–63.
Law, J., et al. (1987). Technology and heterogeneous engineering: the case of Portuguese expansion. *The social construction of technological systems:, New directions in the sociology and history of technology*, *1*, 1–134.
LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.
Lee, K.F. (2017). The real threat of artificial intelligence. The New York Times 24.
Lee, K.F. (2018). AI superpowers: China, Silicon Valley, and the New World Order. Houghton Mifflin Harcourt.
Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14–19.
Madras, D., Creager, E., Pitassi, T., Zemel, R. (2019). Fairness through causal awareness: learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 349–358). New York: ACM.
Maldonado-Torres, N. (2007). On the coloniality of being: contributions to the development of a concept. *Cultural studies*, *21*(2-3), 240–270.
Marda, V., & Narayan, S. (2020). Data in New Delhi's predictive policing system. In *Proceedings of the 2020 conference on fairness, accountability, and transparency,* pp. 317–324.
McClintock, A. (1992). The angel of progress: pitfalls of the term "post-colonialism". Social text (31/32), 84–98.
McDowell, C., & Chinchilla, M.Y. (2016). 30 partnering with communities and institutions. Civic Media: Technology, Design, Practice , 461.
Mezzadra, S., & Neilson, B. (2017). On the multiple frontiers of extraction: excavating contemporary capitalism. *Cultural Studies*, *31*(2-3), 185–204.
Mignolo, W.D. (2007). Introduction: coloniality of power and de-colonial thinking. *Cultural studies*, *21*(2-3), 155–167.
Mignolo, W.D. (2012). *Local histories/global designs: coloniality, subaltern knowledges, and border thinking*. Princeton: Princeton University Press.

Mikesell, L., Bromley, E., Khodyakov, D. (2013). Ethical community-engaged research: a literature review. *American Journal of Public Health*, *103*(12), e7–e14.
Milan, S. (2013). *Social movements and their technologies: wiring social change*. Berlin: Springer.
Milan, S., & Treré, E. (2019). Big data from the south (s): beyond data universalism. *Television & New Media*, *20*(4), 319–335.
Milan, S., & Van der Velden, L. (2016). The alternative epistemologies of data activism. *Digital Culture & Society*, *2*(2), 57–74.
Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220–229). New York: ACM.
Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K. (2018). Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. arXiv:181107867.
Mitchell, T.M. (2006). The discipline of machine learning. Tech. Rep CMU-ML-06-108, Machine Learning Department, Carnegie Mellon University.
Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. In *Thirty-second AAAI conference on artificial intelligence*.
Nandy, A. (1989). *Intimate enemy: loss and recovery of self under colonialism*. Oxford: Oxford University Press Oxford.
Ndlovu-Gatsheni, S.J. (2015). Decoloniality as the future of Africa. *History Compass*, *13*(10), 485–496.
Nelson, A. (2002). Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association*, *94*(8), 666.
Nissenbaum, H. (2001). How computer systems embody values. *Computer*, *34*(3), 120–119.
Noble, S.U. (2018). *Algorithms of oppression: how search engines reinforce racism*. New York: NYU Press.
Nuremberg Military Tribunals (1947). Permissible medical experiments. In *Trials of war criminals before the nuremberg military tribunals under control council law No. 10, vol 2, U.S. government printing office,* pp 181–182.
Nyabola, N. (2018). Digital democracy, analogue politics: how the Internet era is transforming politics in kenya. Zed Books Ltd.
Nyawa, J.M. (2019). The big brother is watching: Huduma Namba a threat to our rights and freedoms. Available at SSRN 3389268.
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.
OECD (2019). OECD principles on artificial intelligence. https://www.oecd.org/going-digital/ai/principles/.
O'Neil, C. (2016). Weapons of math destruction: how big data increases inequality and threatens democracy. Broadway Books.
Pappas, G.F. (2017). The limitations and dangers of decolonial philosophies: lessons from Zapatista Luis Villoro. Radical Philosophy Review.
Pathways for Prosperity (2019). Digital diplomacy: technology governance for developing countries. Pathways for Prosperity Commission on Technology and Inclusive Development. https://pathwayscommission.bsg.ox.ac.uk/sites/default/files/2019-10/Digital-Diplomacy.pdf.
Philip, K., Irani, L., Dourish, P. (2012). Postcolonial computing: a tactical survey. *Science, Technology, & Human Values*, *37*(1), 3–29.
Van de Poel, I., & Kroes, P. (2014). Can technology embody values? In *The moral status of technical artefacts* (pp. 103–124). Berlin: Springer.
Pollard, J., Mcewan, C., Hughes, A. (2011). *Postcolonial economies*. London/New York: Zed Books.
Quijano, A. (2000). Coloniality of power and Eurocentrism in Latin Americas. *International Sociology*, *15*(2), 215–232.
Quijano, A. (2007). Coloniality and modernity/rationality. *Cultural studies*, *21*(2-3), 168–178.
Raji, I.D., & Dobbe, R. (2020). Concrete problems in AI safety, revisited. In *ICLR workshop on ML in the real world*.
Ramose, M.B. (1999). African philosophy through Ubuntu. Mond Books.
Ricaurte, P. (2019). Data epistemologies, the coloniality of power, and resistance. *Television & New Media*, *20*(4), 350–365.

Richardson, R., Schultz, J., Crawford, K. (2019). Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems and justice. New York University Law Review Online, Forthcoming.

Rock, D., & Grant, H. (2016). Why diverse teams are smarter. *Harvard Business Review*, *4*(4), 2–5.

Ronconi, L. (2015). Enforcement and the effective regulation of labor. Tech. rep., IDB Working Paper Series.

Røpke, I. (2001). New technology in everyday life–social processes and environmental impact. *Ecological economics*, *38*(3), 403–422.

Russell S, & Norvig P. (2016). *Artificial intelligence: a modern approach*. New Jersey: Prentice Hall.

Said, E.W. (1993). Culture and imperialism. Vintage.

Sap, M., Card, D., Gabriel, S., Choi, Y., Smith, N.A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics,* pp 1668–1678.

Schroeder, D., Cook Lucas, J., Hirsch, F., Fenet, S., Muthuswamy, V. (2018). *Ethics dumping case studies from north-south research collaborations*. Cham: Springer International Publishing.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency,* (pp. 59–68). New York: ACM.

Sengers, P., Boehner, K., David, S., Kaye, J. (2005). In *Proceedings of the 4th decennial conference on critical computing: between sense and sensibility, pp 49–58*: Reflective design.

Sengers, P., McCarthy, J., Dourish, P. (2006). Reflective HCI: Articulating an agenda for critical practice. In *CHI'06 extended abstracts on Human factors in computing systems* (pp. 1683–1686). New York: ACM.

Senior, E. (2018). *The Caribbean and the medical imagination, 1764-1834: slavery, disease and colonial modernity* Vol. 119. Cambridge: Cambridge University Press.

Shore, N. (2006). Re-conceptualizing the Belmont report: a community-based participatory research perspective. *Journal of Community Practice*, *14*(4), 5–26.

Siddiqui, A.U., & Singh, H.K. (2015). "Aadhar" management system. *IITM Journal of Management and IT*, *6*(1), 40–43.

Sokoloff, N.J., & Pincus, F.L. (2008). Introduction: race, class, gender, and capitalism. Race, Gender & Class, 4–8.

Stark, L. (2019). Facial recognition is the plutonium of AI. XRDS: Crossroads. *The ACM Magazine for Students*, *25*(3), 50–55.

Steiner, C.B. (1994). Technologies of resistance: structural alteration of trade cloth in four societies. Zeitschrift fü,r Ethnologie, pp. 75–94.

Stoler, A.L. (2008). Epistemic politics: ontologies of colonial common sense. In *The philosophical forum*, (Vol. 39 pp. 349–361). New Jersey: Wiley Online Library.

Taylor, E. (2016). Groups and oppression. *Hypatia*, *31*(3), 520–536.

Thatcher, J., O'Sullivan, D., Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: new metaphors for daily data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006.

Thrush, C. (2008). American curiosity: cultures of natural history in the colonial British Atlantic world. *Environmental History*, *13*(3), 573.

Tilley, H. (2014). Conclusion: experimentation in colonial East Africa and beyond. *The International Journal of African Historical Studies*, *47*(3), 495–505.

Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D.CM., Ezer, D., van der Haert, F.C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M., Glasmachers, T., de Wever, W., Teh, Y.W., Khan, M.E., De Winne, R., Schaul, T., Clopath, C. (2020). AI for social good: unlocking the opportunity for positive impact. *Nature Communications*, *11*(1), 1–6. Nature Publishing Group.

Toyama, K. (2015). *Geek heresy: rescuing social change from the cult of technology*. New York: PublicAffairs.

UK National Health Service (2019). Code of conduct for data-driven health and care technology.

UNCTAD. (2013). *Information economy report 2013: The cloud economy and developing countries*. Geneva: United Nations Conference on Trade and Development.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., Jiang, N. (2013). Children in the public benefit system at risk of maltreatment: identification via predictive modeling. *American journal of preventive medicine*, *45*(3), 354–359.
Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S.D., Tegmark, M., Nerini, F.F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*(1), 1–10.
Vitak, J., Shilton, K., Ashktorab, Z. (2016). Beyond the Belmont principles: ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing* (pp. 941–953). New York: ACM.
Wa Thiong'o, N. (1992). Decolonising the mind: the politics of language in African literature. East African Publishers.
Wallerstein, I. (1987). World-systems analysis.
Washington, H.A. (2006). Medical apartheid: the dark history of medical experimentation on Black Americans from colonial times to the present. Doubleday Books.
Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research*. London: Nuffield Foundation.
Winner, L. (1980). Do artifacts have politics? Daedalus, 121–136.
WMA General Assembly (1964). Ethical principles for medical research involving human subjects. World Medical Association Declaration of Helsinki. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566407/pdf/11357217.pdf.
Wong, P.H. (2012). Dao, harmony and personhood: towards a confucian ethics of technology. *Philosophy & technology*, *25*(1), 67–86.
Young, M., Magassa, L., Friedman, B. (2019). Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, *21*(2), 89–103.
Yuan, L. (2018). *How cheap labor drives China's A.I. ambitions*. New York: New York Times. https://www.nytimes.com/2018/11/25/business/china-artificial-intelligence-labeling.html.
Zembylas, M. (2017). Love as ethico-political practice: inventing reparative pedagogies of aimance in "disjointed" times. *Journal of curriculum and pedagogy*, *14*(1), 23–38.
Zuboff, S. (2019). *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. London: Profile Books.

Affiliations

Shakir Mohamed[1] · Marie-Therese Png[2] · William Isaac[1]

Marie-Therese Png
marie-therese.png@oii.ox.ac.uk

William Isaac
williamis@deepmind.com

[1] DeepMind, London, UK

[2] University of Oxford, Oxford, UK