# NEG

## 1nc Digital Democracy

#### US democracy promotion is imperialist, hypocritical and bolsters authoritarian surveillance

Klyman 22

Kevin Klyman researches US-China relations and has written data protection policies adopted by the World Health Organization. “Biden’s Campaign for “Digital Democracy” Is Really a Giveaway to Big Tech”, Jacobin. 6/26/22. <https://jacobin.com/2022/06/us-tech-companies-government-contracts-data-google-facebook-microsoft-amazon/> --JrH

American democracy promotion has been a calamity, to put it lightly. This century alone, the United States has helped violently overthrow the governments of Iraq, Afghanistan, Libya, Haiti, Bolivia, and Honduras, leaving [millions](https://www.bu.edu/pardee/2021/09/02/costs-of-war-project-releases-updated-estimates-of-human-and-financial-costs-of-post-9-11-wars/) [dead](https://web.mit.edu/humancostiraq/) and tens of millions more [displaced](https://watson.brown.edu/costsofwar/files/cow/imce/papers/2020/Displacement_Vine%20et%20al_Costs%20of%20War%202020%2009%2008.pdf), all in the name of democracy. The irony is potent: at home, the United States responded to 2020’s mass protests against police brutality with yet more police brutality, and 2021 began with an attempted coup galvanized by the outgoing president and his political party. Despite the wreckage and the hypocrisy, democracy promotion remains a centerpiece of US foreign policy, and is mobilized as a justification for American goals, whatever they may be. With respect to technology, US policymakers have called for the promotion of “digital democracy” and opposition to “digital authoritarianism” emanating from China. The narrative of democracy triumphing over high-tech dictatorship obscures America’s real goal, which is to prevent Beijing from displacing Washington as the leader in global surveillance and the owner of much of the world’s internet infrastructure. US technology strategy has the same underlying motivation as many other policies that ostensibly aim to promote democracy: opening up markets so that American firms can sell their products abroad. What the tech industry and policymakers have dubbed “digital democracy” is just a recapitulation of US imperialism with respect to the pursuit of global technology dominance. Promoting digital democracy is dangerous for three reasons. First, the United States and other electoral democracies engage in indiscriminate mass surveillance around the world. Second, the US government foists technology policies on poor countries that largely benefit Big Tech. Third, the narrative that innovative American companies are more righteous than their illiberal Chinese counterparts manufactures consent for US-backed surveillance. Voting for the Panopticon “Digital democracy” is by nature defined against digital authoritarianism. The Brookings Institution — which [receives funding](https://www.brookings.edu/wp-content/uploads/2020/04/The-Brookings-Institutions-Contributors-List-Fiscal-Year-2020.pdf) from Google, Facebook, and Amazon — [defines](https://www.brookings.edu/wp-content/uploads/2019/08/FP_20190827_digital_authoritarianism_polyakova_meserole.pdf) digital authoritarianism as “the use of digital information technology by authoritarian regimes to surveil, repress, and manipulate domestic and foreign populations.” This raises the question: What about when democracies use the same tactics? The Information Technology and Innovation Foundation — a think tank [backed](https://itif.org/our-supporters) by Apple, Microsoft, and Uber — [answers](https://itif.org/publications/2021/01/19/us-grand-strategy-global-digital-economy) that democracies, unlike authoritarian states, would never abuse technology: “Authoritarian nations will use technology for authoritarian purposes. Democratic nations will use them for legitimate and civil-liberty-protecting purposes.” This kind of black-and-white thinking creates a self-reinforcing logic: when democratic states and their tech companies engage in surveillance, we can assume that it is for the cause of freedom, but when authoritarian states do it, it is for the purpose of social control. Was the NSA [listening in](https://www.reuters.com/article/us-germany-usa-spying/u-s-spy-agency-tapped-german-chancellery-for-decades-wikileaks-idUSKCN0PI2AD20150709) on Angela Merkel’s phone calls a triumph of representative government? Are YouTube’s efforts to [deplatform](https://www.aljazeera.com/news/2018/1/20/palestinians-fight-facebook-youtube-censorship) Palestinian activists justified by the will of the majority? Is Facebook facilitating genocide in Myanmar and [Ethiopia](https://www.theguardian.com/technology/2021/oct/07/facebooks-role-in-myanmar-and-ethiopia-under-new-scrutiny) “what democracy looks like”? Apparently so. In truth, spying on people and stealing their data is abhorrent regardless of the nature of the government performing or sanctioning that surveillance. Regrettably, surveillance capitalism has become a [driving force](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2594754) in the global economy. Across the world’s democracies, republics, oligarchies, monarchies, theocracies, and dictatorships, accumulating data for profit has become the modus operandi of many companies. US technology strategy has the same underlying motivation as other policies that ostensibly aim to promote democracy: opening up markets so that American firms can sell their products abroad.

#### Surveillance capitalism weaponizes digital technology to control communities and turns democracy promotion – US/Big tech partnership reproduces economic inequality, human rights violation and environmental degradation

Hynes 21

[Mike](file:////insight/search%3fq=Mike%20Hynes" \o "Mike Hynes) Hynes, PhD, is  lecturer in Political Science and Sociology at the National University of Ireland Galway, specializing in environmental sociology, Mobilities and sustainability research. (2021). "Digital Democracy: The Winners and Losers", The Social, Cultural and Environmental Costs of Hyper-Connectivity: Sleeping Through the Revolution, Emerald Publishing Limited, Bingley, pp. 137-153. <https://doi.org/10.1108/978-1-83909-976-220211009> --JrH

The digital ICT revolution promised much for democratic politics in the twenty-first century but so far has delivered little but disruption. The dawn of the internet age was to bring a decisive shift towards the citizen and information was to become free and limitless, and enlightenment and empowerment would follow. But while digital technologies provide us with the opportunity to accumulate quantities of information that one time may not have been possible, big tech and the state remains much better equipped than any private citizen to take full advantage of this opportunity. In many ways, digital technology has been weaponised against the very system it was purported to support and defend and the citizens it was meant to engage, protect and enlighten. Authoritarian regimes across the world have seized upon the opportunities provided by such technology to increase surveillance and control of their people while simultaneously spreading misinformation and confusion, undermining many of the established Western liberal democracies. It would be rather naïve to think that democratic governments are not also regularly using similar digital surveillance technique under various guises and security apparatuses. And all the while big tech is the real big winner. The pioneers of surveillance capitalism Google were emboldened and benefitted from historical events when a national security apparatus, galvanised by the attacks of 11 September 2001, saw the emergent capabilities and the promise of some certainty in how Google’s storage and use of huge stocks of personal data could be used to shadow and predict the behaviour of individuals.[37](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn37)Zuboff believes that the concepts underpinning surveillance capitalism are facilitating the overthrow of the people’s sovereignty and is a prominent force in the perilous drift towards democratic deconsolidation that now threatens Western liberal democracies themselves. And this is a common complaint in the twenty-first century; democracy itself has lost control of corporate power in the form of big tech companies, who use whatever means possible to hoard vast wealth and influence while fuelling inequality, damaging the planet and avoid paying their fair share of taxes.[38](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn36)Today’s big tech behemoths exist in a political culture that has grown accustomed and accommodating to their every need, and Runciman argues, in the United States, this was further cemented by the Supreme Court decision in the Citizens United case of 2010 to grant corporations the same rights to free speech as individual citizens.[39](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn39) The ideals and very notion of liberal democracy are now under constant pressure from many angles, and the traditional hierarchy of power is also under increasing danger. The power of modern corporate power, in the form of big tech, has grown exponentially over the past decade to the point where it now has the wherewithal to undermine how democracy itself operates and not be overly worried about the consequences. A major imperative now for every citizen and democratic nation must be to reassess the inequitable influence of big tech corporate power and the internet, particularly as it relates to our personal data, and to question: who owns and controls such power, and what right do they have to use and misuse our personal data to undermine our key democratic institutions? Democracy must be seen to represent the wishes of the people rather than viewed as a system of corporate tyranny.

#### Alternative is democratic digital collective – solves surveillance capitalism, cultivates communal and accessible response to sociotechno exploitation

Martell 21

Professor Luke Martell has been a Lecturer, Senior Lecturer, Reader, and Professor of Political Sociology at Sussex. He was Head of the Department of Sociology, Chair of the Social and Political Thought Graduate programme, and President of the University and College Union branch. He is now a part-time Teaching Fellow in the Sociology Department. He teaches modules on 'The Death of Socialism?' and 'Alternative Societies'. His book 'Alternative Societies: For a Pluralist Socialism' will be published by Bristol University Press in 2023. “[Surveillance capitalism and digital alternatives](https://www.ideology-theory-practice.org/blog/surveillance-capitalism-and-digital-alternatives)”, 8/11/21. https://www.ideology-theory-practice.org/blog/surveillance-capitalism-and-digital-alternatives --JrH

So, to recap and clarify key points. Oligopoly and the harvesting and selling of our digital lives has become a norm and a new economic sector of capitalism. State responses, to very different degrees, have been to resist monopolization and ensure modest privacy protections or awareness. Individual responses and those of some organizations have been to use software that blocks tracking and aims to maintain privacy and anonymity. But positive as these methods are, they are in part defensive, limited in what they can achieve against high-level attempts at intrusion, and some of these individualise action. Alongside such state and individual processes, we need a more pro-active and collective approach. This includes stronger regulation and breaking up and taking tech into collective ownership. In the sphere of alternatives, it means expanding and strengthening a parallel sphere, decentralised and federated. And alternatives require putting control in the hands of those affected, so collective democracy with inclusive participation. Then oligopolies are challenged and there is a link between those affected and those in control. But alternatives must be made accessible and more easily understandable to the non-techy and beyond the expert, and do not just have to be an alternative but can be a prefigurative basis for spreading to the way the digital and tech world is more widely. This involves supplementing liberal individual privacy and rights approaches, often defensive within the status quo, with collective democracy and control approaches, more proactive and constructive of alternatives[[7]](https://www.ideology-theory-practice.org/blog/surveillance-capitalism-and-digital-alternatives#_edn7). If there is an erosion of capitalism out of such an approach so there will be also to profit incentives in surveillance capitalism. With an extension of collective control not-for-profit, then motivations for surveillance and data capture are reduced. But this must be done through inclusive democratic control (by workers, users and the community) as much as possible rather than the traditional state, as the latter has its own reasons for surveillance. It should be supplemented by a pluralist, decentralised, federated, digital world to counter oligopoly and power. Democratisation that is inclusive globally is also suited to dealing with differences and divides digitally, e.g. by class or across the Global North and Global South. Taken together this approach implies pluralist democratic socialism as well as liberalism, rather than capitalism or the authoritarian state.

## 1nc Governance

#### The aff’s AI governance policy is coloniality that locks the global south into an imbalanced power relation to colonial powers

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 669-671)-qcl

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 a 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 670 S. Mohamed et al. 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.

#### The alternative is decolonial AI – Critical Technical Practice centers decoloniality to produce ethical research strategies, safety, diversity initiatives and activism and disrupt colonial Sociotechnical discourses that enframe AI research & development

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 669-671 pp. 671)-qcl

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 Decolonial AI: Decolonial Theory as Sociotechnical 673 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).

## Links

### Techno Dominance

#### Aff’s Techno hypocrisy obscures violence and economic exploitation that undergirds US global tech hegemony – sale of facial recognition tech, satellite primacy and strongarm data agreements policies

Klyman 22

Kevin Klyman researches US-China relations and has written data protection policies adopted by the World Health Organization. “Biden’s Campaign for “Digital Democracy” Is Really a Giveaway to Big Tech”, Jacobin. 6/26/22. <https://jacobin.com/2022/06/us-tech-companies-government-contracts-data-google-facebook-microsoft-amazon/> --JrH

US policymakers portray China’s technology strategy as uniquely dangerous and despotic, but what they truly fear is that China threatens to displace the United States as the global leader in advanced technologies. Take digital infrastructure as an example. Jake Sullivan, Joe Biden’s national security advisor, has [claimed](https://foreignpolicy.com/2020/05/22/china-superpower-two-paths-global-domination-cold-war/) that China is pursuing “global domination” by building out internet infrastructure in Eurasia and Africa through its Belt and Road Initiative. However, the United States is a far more dominant player in global technology infrastructure. The US government and American firms [own](https://www.ucsusa.org/resources/satellite-database) three thousand satellites, 60 percent of the global total, [including](https://techcrunch.com/2018/12/21/the-gps-wars-have-begun/) the Global Positioning System (GPS), a constellation of thirty satellites that provides location services to billions of devices. US companies design [three-quarters](https://www.semiconductors.org/wp-content/uploads/2021/05/BCG-x-SIA-Strengthening-the-Global-Semiconductor-Value-Chain-April-2021_1.pdf) of all semiconductors, the microelectronics necessary for smartphones and computers. And Google’s Android operating system [powers](https://gs.statcounter.com/os-market-share/mobile/worldwide) 72 percent of the world’s mobile phones — Apple’s iOS accounts for the remaining 28 percent — allowing Google to [leach](https://arstechnica.com/gadgets/2021/03/android-sends-20x-more-data-to-google-than-ios-sends-to-apple-study-says/) personal information from 2.5 billion people round-the-clock. Nanjira Sambuli, one of East Africa’s leading technology policy analysts and a Ford Global Fellow, said in an interview with Jacobin, “For the average user, the interaction they have with Chinese tech is phones. But the phones run Android.” When they bother to acknowledge the reality of US global tech hegemony at all, believers in American exceptionalism contend that even if US-based companies monopolize essential infrastructure and engage in mass surveillance, their actions are qualitatively different from those of Chinese firms, as they are less integrated with the state. In reality, US tech companies are intimately connected with the government. Big Tech receives billions of dollars in corporate welfare each year in the form of [subsidies](https://www.theguardian.com/cities/2018/jul/02/us-cities-and-states-give-big-tech-93bn-in-subsidies-in-five-years-tax-breaks) and [tax breaks](https://fairtaxmark.net/silicon-six-end-the-decade-with-100-billion-tax-shortfall/). The Pentagon and the Department of Homeland Security [pay](https://bigtechsellswar.com/)billions to tech companies in exchange for cloud computing services, secure databases, and augmented reality systems. Besides monetary support, the revolving door between Big Tech and the federal government is [notorious](https://therevolvingdoorproject.org/personnel/), with tech-backed nominees [populating](https://therevolvingdoorproject.org/the-industry-agenda-big-tech/) key regulatory positions. Munira Lokhandwala, director of tech and training at the anti-corruption group Little Sis, told Jacobin, “Virtually every US government department has multiple multiyear contracts with Big Tech.” According to Lokhandwala, “Some agencies are wholly reliant on tech companies. As a government employee, there’s often no way to do your work without outsourcing some of it to one of the five major tech companies.” Chinese firms are rightly [criticized](https://www.euppublishing.com/doi/full/10.3366/ajicl.2022.0393) for [exporting](https://blogs.lse.ac.uk/africaatlse/2021/09/09/dont-blame-china-for-rise-of-digital-authoritarianism-africa-surveillance-capitalism/) surveillance technology abroad to autocratic governments such as Uganda, Zambia, and Kazakhstan. Similarly, US companies [should](https://www.cnet.com/tech/mobile/clearview-ai-probed-over-facial-recognition-sales-to-foreign-governments/) be condemned for [selling](https://www.washingtonpost.com/outlook/2019/01/17/how-us-surveillance-technology-is-propping-up-authoritarian-regimes/) facial recognition technology to Saudi Arabia, [Egypt](https://www.zawya.com/en/press-release/honeywell-to-deploy-world-class-public-safety-and-security-infrastructure-for-egypts-new-smart-city-f8th3lyy), [Israel](https://www.theverge.com/2020/3/27/21197577/microsoft-facial-recognition-investing-divest-anyvision-controversy), and the United Arab Emirates, helping these totalitarian states crack down on dissidents. In 2020, Belarus [used](https://www.bloomberg.com/news/articles/2020-09-11/sandvine-use-to-block-belarus-internet-rankles-staff-lawmakers) equipment provided by the American company Sandvine to block millions of websites and censor news and social media in the midst of mass protests following Alexander Lukashenko’s [sham](https://www.economist.com/leaders/2020/08/13/belaruss-election-was-a-sham-the-wests-response-has-been-feeble) reelection. After several outlets reported its involvement, Sandvine quietly [ended](https://www.bloomberg.com/news/articles/2020-09-15/sandvine-says-it-will-no-longer-sell-its-products-in-belarus) its partnership with the Belarusian government, though it [continued](https://www.bloomberg.com/news/articles/2022-06-03/sandvine-pulls-back-from-russia-as-us-eu-tighten-control-on-technology-it-sells) similar partnerships in Jordan and Russia. Francisco Partners, which owns Sandvine, was later “[crowned](https://www.franciscopartners.com/news/francisco-partners-crowned-2020-s-top-money-maker) 2020’s top money maker” among tech-focused private equity firms.

### Data Privacy

#### US assault on data protection/privacy bolsters techno hegemony and reflects a larger project of public/private collusion to deny fundamental human right of privacy to millions – turns democracy promotion

Klyman 22

Kevin Klyman researches US-China relations and has written data protection policies adopted by the World Health Organization. “Biden’s Campaign for “Digital Democracy” Is Really a Giveaway to Big Tech”, Jacobin. 6/26/22. <https://jacobin.com/2022/06/us-tech-companies-government-contracts-data-google-facebook-microsoft-amazon/> --JrH

Freedom to Assemble Our Computer Overlords The next pillar of America’s “pro-democracy” digital strategy is to weaken other countries’ data protection and privacy laws. The [dominant view](https://nationalsecurity.gmu.edu/2019/05/combating-digital-authoritarianism-u-s-alternative-needed-to-counter-data-localization-and-government-control/) in some circles of “the [Blob](https://jacobin.com/2021/05/biden-administration-foreign-policy-china-forever-wars)” is that data protection can exacerbate authoritarianism because dictators use such laws to ensure that the “government is the main arbiter and moderator of data.” The US International Trade Commission — a federal agency tasked with adjudicating international trade disputes — staunchly [opposes](https://www.usitc.gov/publications/332/pub4415.pdf) strong data protection laws, suggesting they tend “to cause substantial damage to consumer trust in the Internet; to erode business opportunities for data-related innovations, for example, in the areas of analytics and Big Data; and to raise costs for businesses complying with multiple divergent standards.” US technology giants wholeheartedly agree that private companies, not governments, should be the main arbiters and moderators of data. Big Tech companies have accordingly become the leading opponents of robust data protection and privacy laws across the world. Leaked Facebook documents revealed that it was able [block](https://www.theguardian.com/technology/2019/mar/02/facebook-global-lobbying-campaign-against-data-privacy-laws-investment) privacy legislation in Malaysia by promising the government it would increase investment. In Kenya, Google and IBM are at the [forefront](https://www.ids.ac.uk/opinions/lobbying-for-digital-dominance-in-africa/) of undercutting Kenya’s data protection law by lobbying for looser restrictions on the transfer of personal data. In Europe, where the EU is [considering](https://www.wired.com/story/artificial-intelligence-regulation-european-union/) harsher penalties for algorithmic harm, Apple, Google, Facebook, and Microsoft are the continent’s [largest lobbyists](https://corporateeurope.org/sites/default/files/2021-08/The%20lobby%20network%20-%20Big%20Tech%27s%20web%20of%20influence%20in%20the%20EU.pdf). “America’s ‘more democratic’ alternative to allowing countries to manage their own data is to coerce them into allowing US multinationals to host their citizens’ personal data in the US.” Companies have brazenly violated the law to establish monopolies overseas. In 2019, the Securities and Exchange Commission [fined](https://www.wsj.com/articles/microsoft-to-pay-25-million-to-settle-foreign-bribery-probe-11563811097) Microsoft $25 million for bribing the governments of Hungary, Saudi Arabia, Thailand, and Turkey to use Microsoft’s services. A Microsoft whistleblower subsequently [revealed](https://www.lioness.co/post/microsoft-is-using-illegal-bribes-in-the-middle-east-and-africa-why-is-the-sec-turning-a-blind-eye) in March 2022 that the company pays $200 million annually in bribes and kickbacks to secure favorable government contracts in the Middle East and North Africa, including in Ghana, Nigeria, Zimbabwe, and Qatar. Although the terms of Microsoft’s government contracts are not public, it is likely that Microsoft’s contracts contain few privacy provisions as they “cement its monopoly on the continent,” in the whistleblower’s words. Sambuli, the Ford Global Fellow, told Jacobin, “What’s happening behind the scenes is what we don’t know. We don’t know what those companies have negotiated. We don’t know which parts of us have been sold.” Despite Big Tech’s protestations, data protection and privacy laws are usually motivated by their clear-cut benefits, not tyrannical ploys for government supervision. Privacy is a [fundamental](https://www.hrw.org/world-report/2014/essays/privacy-in-age-of-surveillance) human right. [Half](https://unctad.org/news/least-developed-countries-still-lag-behind-cyberlaw-reforms) of the world’s poorest countries have no data protection or privacy legislation, leaving their citizens at risk of having their personal information stolen or [sold](https://privacyinternational.org/long-read/3390/2020-crucial-year-fight-data-protection-africa) to third parties. Before the [adoption](https://www.insideprivacy.com/data-privacy/tech-regulation-in-africa-recently-enacted-data-protection-laws/) of South Africa’s Protection of Personal Information Act in 2020, the personal information of 30 million South Africans was hacked, a figure representing more than [half](https://www.bbc.com/news/world-africa-41696703) the country. Data protection laws boost countries’ economies as well, helping [eliminate](https://www.cgdev.org/publication/why-data-protection-matters-development-case-strengthening-inclusion-and) legal ambiguity and reassuring companies that, if they expand in a country, their data will be safe and they will not face arbitrary litigation. But even more importantly, data privacy ensures that people feel safe enough to organize and express their political beliefs. In other words, it is fundamental to democracy — exposing the absurdity of the United States’ opposition in the name of democracy promotion.

#### US digital authoritarian strategy reinforces economic imbalances between countries and magnifies inequality in the poorest countries – punitive approaches to multinational service providers generate millions in digital service taxes Klyman 22

Kevin Klyman researches US-China relations and has written data protection policies adopted by the World Health Organization. “Biden’s Campaign for “Digital Democracy” Is Really a Giveaway to Big Tech”, Jacobin. 6/26/22. <https://jacobin.com/2022/06/us-tech-companies-government-contracts-data-google-facebook-microsoft-amazon/> --JrH

Another dangerous aspect of the United States’ digital strategy is its punitive approach toward countries that levy taxes on multinationals that provide digital services. Digital services taxes are often cited as a core component of digital authoritarianism in [Russia](https://www.american.edu/sis/centers/security-technology/russian-cyber-sovereignty.cfm), since they weaken foreign firms and allow the Kremlin to nurture more pliable domestic providers of digital services. China, by contrast, has [no need](https://www.tandfonline.com/doi/full/10.1080/13602381.2022.2012992) for such taxes, as it has already driven out most major foreign tech companies through internet censorship and invasive data-sharing requirements. Digital services taxes can have downsides, of course. If they do not narrowly target foreign firms, such taxes can [hamper](https://www.economist.com/middle-east-and-africa/2017/11/09/how-the-taxman-slows-the-spread-of-technology-in-africa) the domestic technology ecosystem. Taxes that target firms’ transactions instead of their profits are likely to [increase costs](https://www.bbc.com/news/world-africa-61248366) for consumers, making it more expensive to go online. But they have benefits too. Digital services taxes secure significant revenue for countries with limited tax bases. Malaysia [raised](https://www.vertexinc.com/resources/resource-library/malaysias-tax-digital-services-raises-over-rm400-million) $126 million from its digital services tax in just one year, equal to 1.5 percent of its [overall](https://www.ceicdata.com/en/indicator/malaysia/tax-revenue) tax revenue. Smaller economies have been just as successful: Kenya [raised](https://techcabal.com/2021/01/22/kenyan-government-out-to-raise-45million-digital-tax-revenue-before-july/) $45 million and Ecuador [raised](https://www.cpapracticeadvisor.com/tax-compliance/news/21242174/latin-american-countries-lead-world-in-taxing-digital-services) $20 million in the first year of their digital services taxes. Over time, these taxes [could generate](https://itif.org/publications/2019/05/13/digital-services-taxes-bad-idea-whose-time-should-never-come) up to 3 percent of a country’s tax revenue simply by levying fees on a few multinationals. Logan Wort, the African Tax Administration Forum’s executive secretary, told Jacobin that this fundraising potential demonstrates “the importance of countries taking such approaches to enhance their tax bases.”These revenues are especially important in the context of the global economic crisis wrought by the pandemic, which has [caused](https://www.barrons.com/news/ngos-sound-alarm-over-pandemic-induced-budget-cuts-in-w-africa-01634205907) governments in the Global South to [slash](https://www.worldbank.org/en/news/press-release/2021/02/22/two-thirds-of-poorer-countries-are-cutting-education-budgets-due-to-covid-19) their budgets. Without digital services taxes, governments around the world would be [much more likely](https://www.forbes.com/sites/taxnotes/2021/03/22/digital-services-taxes-may-be-difficult-to-remove/?sh=2e87ad266e12) to default on their loans from the International Monetary Fund, increasing the risk that creditors will force governments to cut costs by eliminating basic services. The United States has launched a full-scale assault on digital services taxes. In June 2021, the Biden administration [imposed](https://ustr.gov/about-us/policy-offices/press-office/press-releases/2021/june/ustr-announces-and-immediately-suspends-tariffs-section-301-digital-services-taxes-investigations) a 25 percent tariff on a wide range of goods from India, Turkey, Spain, and Britain in response to their digital services taxes. US trade representative Katherine Tai said that the government would not enforce the tariffs for six months, allowing trade negotiations to continue while the threat of harsh economic reprisal loomed. In October 2021, the [corporate media](https://www.washingtonpost.com/us-policy/2021/10/30/biden-g20-global-minimum-tax/) celebrated Treasury secretary Janet Yellen’s “groundbreaking” negotiation of a global minimum tax for corporations. The agreement, which sets the minimum tax rate on corporations at 15 percent [rather](https://www.jacobinmag.com/2021/10/global-tax-system-oecd-corporate-tax-rate-avoidance-minimum-pandora-papers-negotiations-capital-multinationals) than the international community’s original goal of 25 percent, has a gaping loophole: it [requires](https://www.oecd.org/tax/beps/statement-on-a-two-pillar-solution-to-address-the-tax-challenges-arising-from-the-digitalisation-of-the-economy-october-2021.pdf) all 140 signatories to repeal their digital services taxes and not introduce similar taxes in the future. Many of the world’s poorest countries refused to sign the deal, since they wanted to retain or implement digital services taxes to expand their tax bases. This has [sparked fears](https://www.law360.com/articles/1472064) that the United States will impose steep tariffs on “half of Africa.” In an interview with Jacobin, Wort warned that such a move would have “broader economic impacts on countries with economic ties to the US” and that countries that have not signed the agreement face “significant risk and continued loss of revenue” if they wait to adopt digital services taxes. The United States is ‘waging war’ on countries that impose nontariff barriers to digital trade — even democratic ones. Though these coercive tactics are a [boon](https://prospect.org/economy/ipef-bidens-successor-to-tpp-is-boon-for-big-tech/) for Big Tech, they have little tangible payoff for the United States. The US International Trade Commission [found](https://www.usitc.gov/publications/332/pub4485.pdf) that eliminating all global digital trade barriers would boost US GDP by just 0.1 percent. Nevertheless, the United States is “[waging war](https://www.wired.com/story/the-us-is-waging-war-on-digital-trade-barriers/)” on countries that impose nontariff barriers to digital trade — even democratic ones. Samantha Power, the head of US Agency for International Development, announced at Biden’s Summit for Democracy that her agency would [spend](https://www.usaid.gov/news-information/speeches/dec-10-2021-administrator-power-summit-democracy-event-countering-digital-authoritarianism) an additional $20 million on this variety of lobbying. “We’ll use these funds to help partner nations align their rules governing the use of technology with democratic principles and respect for human rights,” she said. In April, the Biden administration codified its corporatist stance at the State Department’s signing ceremony for the Declaration for the Future of the Internet. The declaration, which was signed by sixty other countries, discourages countries from data localization, [suggesting](https://www.state.gov/wp-content/uploads/2022/04/Declaration-for-the-Future-for-the-Internet.pdf) they “realize the benefits of data free flows [sic].” A leaked [draft](https://www.politico.com/f/?id=0000017c-e71b-d8e1-a57c-efffa3810004) of the initiative’s aims clarifies that the Biden administration hopes to counter the “alternative vision of the Internet as a tool of State control promoted by authoritarian powers such as China and Russia.”

### Governance

#### The aff’s AI governance policy is coloniality that locks the global south into an imbalanced power relation to colonial powers

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 669-671)-qcl

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 a 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 670 S. Mohamed et al. 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

### Development

#### AI and Co-development is a strategy to address CDSs, they are tools for social development – failure of paternalism, technological solutionism, and predatory inclusion

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 669-671 pp. 671)-qcl

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; Tomasev et al. ˇ 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.

#### AI development causes environmental collapse and labor exploitation – while AI itself is already environmentally harmful through infrastructure networks that require unethical mining operations and endless pollution regimes – it also causes a rebound effect in which its proliferation multiplies its own impacts since its increased capabilities are utilized by businesses that are environmentally harmful

Hagendorf 21 (Thilo, AI researcher at the University of Tübingen. 11-25-21, “Blind Spots in AI Ethics,” Orcid, https://doi.org/10.1007/s43681-021-00122-8)-qcl

Ecosystem services do not only have an unprecedented monetary value for humans [164], but they are the very reason for the possibility of human life on earth. Over-exploitation of ecosystems harms future generations [165], poor and already underprivileged people [166], animals [167], and many more. Whereas the building industry, agroindustry, or transport industry seem to be main drivers of ecosystem destruction and climate change, the information and communication industry also play a tangential role. This holds AI and Ethics 1 3 especially true for the AI field. The term “AI” bears a linguistic similarity to the term “cloud”. The notion of cloud computing suggests that it lacks materiality, that it is invisible and placeless, that data are “stored in the troposphere” [168], where in fact, big data is anything but transcendent or amorphous. It is grounded in fragile physical infrastructures, cables, hardware, routers, server buildings, power grids, cooling systems, satellites, etc., all of whom require natural resources. A similar situation unfolds with AI. The term “artificial intelligence” again suggests something immaterial, a mental quality that has seemingly no physical implications. This could not be farther from the truth. To appreciate this, one must switch from a data level, where the real material complexities of AI systems are far out of sight, to an infrastructural level and to the complete supply chain. Here, and frst of all, it becomes visible that the cloud and AI systems are intrinsically intertwined. The cloud builds the necessary condition of AI, and the material implications of the cloud cling onto AI, too. These implications are far-reaching and manifest themselves in global networks of cable infrastructures, labor division, logistics, distribution, and manifold externalities. These networks comprise lithium, tin, cobalt and other mines that deliver essential minerals for electronic components or batteries that are part of every digital mobile device, smelters and refiners that produce acidic, radioactive, and otherwise harmful waste products, storage systems and warehouses for logistics and transportation operations, energy and water hungry data centers, afliated cooling systems, diesel powered generators for backup purposes in cases of blackouts, data annotation factories, collection operations for toxic electronic waste consisting of technical devices with a lifespan of a few years, and many more [2, 169]. All these mining, shipping, manufacturing, and garbage incineration operations are heavily destructive, have a high burden on ecosystems, and come at the cost of human lives, child labor, wildlife populations, natural habitats, toxins in the ground, water, and air, public health, political instability and tensions and low wage labor markets. The material conditions that allow AI usage, especially rare earth elements or “confict minerals”, are also triggers for military operations, violence, murder, and migration that surround the already brutal and slavery-like industry of mining [170]. AI systems are not just demanding in terms of material resources, they also require a lot of energy. Electronic machines, in contrast to combustion engines, can in principle be used sustainably by consuming electricity from renewable energy sources. In practice, however, in many countries, only small proportions of electricity are renewable [171]. Accordingly, powering the computational resources that are required to collect large amounts of training data and to train, test, and apply large AI models comes with a signifcant carbon footprint [172]. Strubell et al. [173] conducted a life cycle assessment of several large AI models and found out that they can emit around three hundred thousand kilograms of carbon dioxide equivalent. The reason for this lies in the many ways machine learning methods go along with a “bigger is better” approach which prioritizes accuracy over efciency [174] with costly trial and error processes which span from practitioners intuitively setting up model parameters all to neural architecture search and other tuning and automated optimization processes. Ultimately, the information and communication industry, which incorporates the AI feld, has a carbon footprint that is bigger than that of the aviation industry [175]. But while there is fight shame, there is no such remorse for AI use, although some AI ethics researchers have tentatively started to develop a critical perspective on the role that AI has in contributing to climate change [174, 176, 177]. Much of this is dependent on where training servers are located, which energy grid is used, how long models are trained, and what hardware accelerators are in use [178]. However, even under perfect conditions where only renewable energy sources are used, it seems likely that AI remains a polluting technology in many industry sectors due to the business purposes it is utilized for. On one hand, AI technologies are heralded as technical solutions to the climate crisis by helping to develop lowemission infrastructures, operate smart grids, help foster sustainable consumption and production, etc. [21, 179]. On the other hand, AI technologies are used to buttress industries and business models that are environmentally harmful—let alone rebound efects in industries that are deemed to be sustainable [180]. In this regard, the tip of the iceberg is the collaboration between the largest AI companies and the fossil fuel industry [181, 182] which does not only comprise the optimization of oil and gas extraction, but goes so far as to actively support climate change deniers [183].

#### Effect of AI development on countries in the Global South

**Lee 17** (Kai-Fu Lee Taiwanese computer scientist, businessman, and writer, 6-24-2017, accessed on 6-24-2022, The New York Times, "Opinion | The Real Threat of Artificial Intelligence", https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html)-qcl

BEIJING — What worries you about the coming world of artificial intelligence? Too often the answer to this question resembles the plot of a sci-fi thriller. People worry that developments in A.I. will bring about the “singularity” — that point in history when A.I. surpasses human intelligence, leading to an unimaginable revolution in human affairs. Or they wonder whether instead of our controlling artificial intelligence, it will control us, turning us, in effect, into cyborgs. These are interesting issues to contemplate, but they are not pressing. They concern situations that may not arise for hundreds of years, if ever. At the moment, there is no known path from our best A.I. tools (like the Google computer program that recently beat the world’s best player of the game of Go) to “general” A.I. — self-aware computer programs that can engage in common-sense reasoning, attain knowledge in multiple domains, feel, express and understand emotions and so on. This doesn’t mean we have nothing to worry about. On the contrary, the A.I. products that now exist are improving faster than most people realize and promise to radically transform our world, not always for the better. They are only tools, not a competing form of intelligence. But they will reshape what work means and how wealth is created, leading to unprecedented economic inequalities and even altering the global balance of power. It is imperative that we turn our attention to these imminent challenges. What is artificial intelligence today? Roughly speaking, it’s technology that takes in huge amounts of information from a specific domain (say, loan repayment histories) and uses it to make a decision in a specific case (whether to give an individual a loan) in the service of a specified goal (maximizing profits for the lender). Think of a spreadsheet on steroids, trained on big data. These tools can outperform human beings at a given task. This kind of A.I. is spreading to thousands of domains (not just loans), and as it does, it will eliminate many jobs. Bank tellers, customer service representatives, telemarketers, stock and bond traders, even paralegals and radiologists will gradually be replaced by such software. Over time this technology will come to control semiautonomous and autonomous hardware like self-driving cars and robots, displacing factory workers, construction workers, drivers, delivery workers and many others. Unlike the Industrial Revolution and the computer revolution, the A.I. revolution is not taking certain jobs (artisans, personal assistants who use paper and typewriters) and replacing them with other jobs (assembly-line workers, personal assistants conversant with computers). Instead, it is poised to bring about a wide-scale decimation of jobs — mostly lower-paying jobs, but some higher-paying ones, too. This transformation will result in enormous profits for the companies that develop A.I., as well as for the companies that adopt it. Imagine how much money a company like Uber would make if it used only robot drivers. Imagine the profits if Apple could manufacture its products without human labor. Imagine the gains to a loan company that could issue 30 million loans a year with virtually no human involvement. (As it happens, my venture capital firm has invested in just such a loan company.) We are thus facing two developments that do not sit easily together: enormous wealth concentrated in relatively few hands and enormous numbers of people out of work. What is to be done? Part of the answer will involve educating or retraining people in tasks A.I. tools aren’t good at. Artificial intelligence is poorly suited for jobs involving creativity, planning and “cross-domain” thinking — for example, the work of a trial lawyer. But these skills are typically required by high-paying jobs that may be hard to retrain displaced workers to do. More promising are lower-paying jobs involving the “people skills” that A.I. lacks: social workers, bartenders, concierges — professions requiring nuanced human interaction. But here, too, there is a problem: How many bartenders does a society really need? The solution to the problem of mass unemployment, I suspect, will involve “service jobs of love.” These are jobs that A.I. cannot do, that society needs and that give people a sense of purpose. Examples include accompanying an older person to visit a doctor, mentoring at an orphanage and serving as a sponsor at Alcoholics Anonymous — or, potentially soon, Virtual Reality Anonymous (for those addicted to their parallel lives in computer-generated simulations). The volunteer service jobs of today, in other words, may turn into the real jobs of the future. Other volunteer jobs may be higher-paying and professional, such as compassionate medical service providers who serve as the “human interface” for A.I. programs that diagnose cancer. In all cases, people will be able to choose to work fewer hours than they do now. Who will pay for these jobs? Here is where the enormous wealth concentrated in relatively few hands comes in. It strikes me as unavoidable that large chunks of the money created by A.I. will have to be transferred to those whose jobs have been displaced. This seems feasible only through Keynesian policies of increased government spending, presumably raised through taxation on wealthy companies. As for what form that social welfare would take, I would argue for a conditional universal basic income: welfare offered to those who have a financial need, on the condition they either show an effort to receive training that would make them employable or commit to a certain number of hours of “service of love” voluntarism. To fund this, tax rates will have to be high. The government will not only have to subsidize most people’s lives and work; it will also have to compensate for the loss of individual tax revenue previously collected from employed individuals. This leads to the final and perhaps most consequential challenge of A.I. The Keynesian approach I have sketched out may be feasible in the United States and China, which will have enough successful A.I. businesses to fund welfare initiatives via taxes. But what about other countries? They face two insurmountable problems. First, most of the money being made from artificial intelligence will go to the United States and China. A.I. is an industry in which strength begets strength: The more data you have, the better your product; the better your product, the more data you can collect; the more data you can collect, the more talent you can attract; the more talent you can attract, the better your product. It’s a virtuous circle, and the United States and China have already amassed the talent, market share and data to set it in motion. For example, the Chinese speech-recognition company iFlytek and several Chinese face-recognition companies such as Megvii and SenseTime have become industry leaders, as measured by market capitalization. The United States is spearheading the development of autonomous vehicles, led by companies like Google, Tesla and Uber. As for the consumer internet market, seven American or Chinese companies — Google, Facebook, Microsoft, Amazon, Baidu, Alibaba and Tencent — are making extensive use of A.I. and expanding operations to other countries, essentially owning those A.I. markets. It seems American businesses will dominate in developed markets and some developing markets, while Chinese companies will win in most developing markets. The other challenge for many countries that are not China or the United States is that their populations are increasing, especially in the developing world. While a large, growing population can be an economic asset (as in China and India in recent decades), in the age of A.I. it will be an economic liability because it will comprise mostly displaced workers, not productive ones. So if most countries will not be able to tax ultra-profitable A.I. companies to subsidize their workers, what options will they have? I foresee only one: Unless they wish to plunge their people into poverty, they will be forced to negotiate with whichever country supplies most of their A.I. software — China or the United States — to essentially become that country’s economic dependent, taking in welfare subsidies in exchange for letting the “parent” nation’s A.I. companies continue to profit from the dependent country’s users. Such economic arrangements would reshape today’s geopolitical alliances. One way or another, we are going to have to start thinking about how to minimize the looming A.I.-fueled gap between the haves and the have-nots, both within and between nations. Or to put the matter more optimistically: A.I. is presenting us with an opportunity to rethink economic inequality on a global scale. These challenges are too far-ranging in their effects for any nation to isolate itself from the rest of the world.

### Machine learning

#### Machine learning techniques create better AI – reward hacking, epistemological limitations, and distributions problems

Heather M**. Roff** , 19 July 20**22,** Senior Research Analyst, Johns Hopkins Applied Physics Laboratory Fellow, Foreign Policy, Brookings Institution Associate Research Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge <https://arxiv.org/pdf/2008.07321.pdf> - Maren Lien

We want artificial intelligence (AI) to be beneficial.1 This is the grounding assumption of most of the attitudes towards AI research. We want AI to be “good” for humanity. We want it to help, not hinder, humans. Yet what exactly this entails **in theory and in practice** is not immediately apparent. Theoretically, this declarative statement subtly implies a commitment to a consequentialist ethics. Practically, some of the more promising **machine learning techniques** to create a robust AI, and perhaps even an artificial general intelligence (AGI) also **commit** one to a **form of utilitarianism**. In both dimensions, the logic of the beneficial AI movement may not in fact create “beneficial AI” in either narrow applications or in the form of AGI if the ethical assumptions are not made explicit and clear. Additionally, as it is likely that reinforcement learning (RL) will be an important technique for machine learning in this area, it is also important to interrogate how RL smuggles in a particular type of consequentialist reasoning into the AI: particularly, a brute form of **hedonistic act utilitarianism**. Since the **mathematical logic** commits one to a **maximization** function, the result is that an AI will inevitably be seeking **more and more rewards**. We have two conclusions that arise from this. First, is that if one believes that a beneficial AI is an ethical AI, then one is committed to a framework that posits ‘benefit’ is tantamount to the greatest good for the greatest number Second, if the AI relies on RL, then the way it reasons about itself, the environment, and other agents, will be through an act utilitarian morality. This proposition may, or may not, in fact be actually beneficial for humanity, (at least for the majority of reasons cited against utilitarianism over the past three hundred years). Indeed, as I will attempt to show here, much of the philosophical insights about the **deficiencies** of utilitarianism apply **directly** to many of the currently cited concrete problems in **AI** safety, such as specifying the wrong objective function, avoiding negative externalities, reward hacking, distribution problems, epistemological limitations, and distributional shift2 . The paper is organized into four sections. The first section lays out how RL could be seen as an implementation of hedonistic act utilitarianism. Section two discusses the various kinds of utilitarianism, as well as many of its well-cited criticisms, particularly how most of its forms ultimately collapse back into a brute act utilitarian evaluation. Moreover, this section argues that the two ways that utilitarianism is often described, as either an evaluative theory of right or as a decision-procedure for individual action, is the same description of how machine learning researchers are approaching the problem of creating beneficial AI. This presents us with a useful intervention point, both within policy but also within research communities to assess whether this is a solvable problem for AI. The third section argues that many of the problems noted in classic arguments about utilitarianism **manifest** themselves in concrete **AI** research safety problems, especially in **reward hacking**. Finally, section four argues that it may be useful for AI researchers to look towards philosophers to see how they have previously attempted to solve such problems. In particular, it may be more simple to look to various constraints on an AI’s action, or as Mackie notes as “device[s] for countering such specific evils” that may result.

#### Reinforced Learners (RL) rewards function logic mirrors classic utilitarian framing

Heather M**. Roff** , 19 July 20**22,** Senior Research Analyst, Johns Hopkins Applied Physics Laboratory Fellow, Foreign Policy, Brookings Institution Associate Research Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge <https://arxiv.org/pdf/2008.07321.pdf> - Maren Lien

While RL is premised on a feedback loop, where an agent tries a particular action or behavior to reach a (preselected/identified) goal and receives a reward signal from the environment providing information about whether it is acting correctly, the environment does not fundamentally change. The environment may produce varying signals to the agent, but the environment as such, does not ontologically change, and so this is not a co-constituting ontology. The feedback loop only goes in one direction, not two. Thus, for the RL agent, a goal remains something external or exogenous to it, and the agent directs its actions towards it. It cannot modify or change its goal. This is important because as AIs become more powerful, the model for learning may need to change. RL agents may be able to reshape not only their environments, but rewards as well. An example may be helpful here. An agent playing chess receives various signals from its environment (such as another agent’s move on the board). The signal tells the agent which actions are not only permissible to take, but which actions may preferable or best to take, given the available knowledge about the game of chess at the time, with the goal of winning the game. This is a feedback loop. However, it is not a co-constituting agent-structure problem. The agent playing chess cannot, once its adversary moves to e5, claim that the board itself changed and now is a 9x9 matrix. The board is the structure; the agents move within it. The conceptual terrain is important here to understanding how the logic of reinforcement learning and the billiard ball model of learning work to produce a particular kind of utility maximizing agent: a hedonistic act utilitarian. While one may object, and claim that the RL utility maximizing agent is a different kind of “utilitarian” than that of the moral philosophers, this is in fact, not wholly true. One is merely the mathematical representation of the logic of the other, and indeed implicitly carries in certain assumptions about broader normative aims. For example, as Sutton and Barto (1998) explain with regards to reinforcement learning, Rewards are in a sense primary, whereas values, as predictions of rewards, are secondary. Without rewards there could be no values, and the only purpose of estimating values is to achieve more reward. Nevertheless, it is values with which we are more concerned when making and evaluating decisions. Action choices are made based on value judgments. 7 Here, a value judgment is a judgment about what sorts of behaviors or actions contain “worth”. But what “worth” entails, is actually about what is “good” all things considered for the decision or task at hand. 8 It is, therefore a teleological or goal-directed action. Moreover, if one desires to avoid circularity, then one cannot define “good” or “value” in terms of itself, so one must find some other thing to use as a signal or content for the “good.” In the RL scheme, the signal is the reward function, and this is used to estimate or “judge” which actions or states will yield the long term best act/choice/decision; that is, what will most likely yield the “value” function. Yet if we grant that this conceptualization of RL as true, then we must also grant that it maps — almost identically — to the logic of the moral theory of utilitarianism as well. Utilitarianism is a consequentialist moral theory that argues that normative properties are determined by some sort of end state (or goal). In easy terms, the best consequences — here operationalized as the most utility — determine the morally right act. Utility can be defined in any number of ways, as I will explain in the next section, however it is most often associated with an aggregate welfare account, usually seen in shorthand as “the greatest good for the greatest number.” One ought to immediately see that RL is very much like utilitarianism because both the RL agent and the utilitarian moral agent seek to determine some present action by a judgment about maximizing the value — or good — of some future state/goal/consequence. Additionally, utilitarianism also seeks coherence and avoiding circularity. Thus, one cannot define the good in terms of itself, but must instead define it in some other (nonmoral) term. Classically, this is done through a reward signal (pleasure or pain).9

### Military Application (general)

#### AI military application is riddled with error—brings error for miscalc and requires human oversight.

**Atherson ’22** (Kelsey Atherton is a military technology journalist based in Albuquerque, New Mexico. His reporting has appeared in Popular Science, Breaking Defense, and The New York Times, “Understand the Errors Introduced by Military AI Application,” <https://www.brookings.edu/techstream/understanding-the-errors-introduced-by-military-ai-applications/>)- lf

On March 22, 2003, two days into the U.S.-led invasion of Iraq, American troops fired a Patriot interceptor missile at what they assumed was an Iraqi anti-radiation missile designed to destroy air-defense systems. Acting on the recommendation of their computer-powered weapon, the Americans fired in self-defense, thinking they were shooting down a missile coming to destroy their outpost. What the Patriot missile system had identified as an incoming missile, was in fact a UK Tornado fighter jet, and when the Patriot struck the aircraft, it killed two crew on board instantly. The deaths were the first losses suffered by the Royal Air Force in the war and the tragic result of friendly fire. A subsequent RAF Board of Inquiry investigation concluded that the shoot-down was the result of a combination of factors: how the Patriot missile classified targets, rules for firing the missiles, autonomous operation of Patriot missile batteries, and several other technical and procedural factors, like the Tornado not broadcasting its “friend or foe” identifier at the time of the friendly fire. The destruction of Tornado ZG710, the report concluded, represented a tragic error enabled by the missile’s computer routines. The shoot-down of the Tornado happened nearly 20 years ago, but it offers an insight into how AI-enabled systems or automated tools on the battlefield will affect the kinds of errors that happen in war. Today, human decisionmaking is shifting toward machines. With this shift comes the potential to reduce human error, but also to introduce new and novel types of mistakes. Where humans might have once misidentified a civilian as a combatant, computers are expected to step in and provide more accurate judgment. Across a range of military functions, from the movement of autonomous planes and cars to identifying tanks on a battlefield, computers are expected to provide quick, accurate decisions. But the embrace of AI in military applications also comes with immense risk. New systems introduce the possibility of new types of error, and understanding how autonomous machines will fail is important when crafting policy for buying and overseeing this new generation of autonomous weapons. What went wrong in 2003 The Patriot missile began development in the 1960s, when the U.S. Army sought a means to reliably shoot down enemy airplanes. Later, the missile would gain the ability to also intercept other missiles, and as the roles assigned to the missile expanded, its autonomous capabilities increased. Patriot missile batteries use a phased array radar to detect and identify targets. This information is then fed into a computer control station to manage how the missiles are launched in response. Once fired, the missiles fly toward an identified intercept point calculated before firing, directions that can be altered by sending updated sensor readings over radio signal to the fired missile. As it approaches for impact, the missile’s own radar tracks the target. Raytheon, which manufactures the Patriot, has described the system as having “automated operations” with “man-in-the-loop (human) override” capabilities—technology that allows the weapon to quickly engage targets with the necessary speed to carry out its missile defense mission. Automation is a compelling feature for an anti-air and, especially, for an anti-missile system. The calculations involved in shooting down aircraft and missiles are hard and require immediate translation of sensor information. Both interceptors and targets are traveling exceptionally fast. It’s the kind of task in which the involvement of a human introduces lag, slows down the process, and makes it less likely a missile is going to successfully shoot down an incoming projectile or aircraft. But human operators also serve an essential role: preventing accidental, incorrect shootdowns. And this requires a balance between human and machine decisionmaking that is difficult to achieve. When the Pentagon investigated the causes of the Tornado shootdown, as well as two other incidents of friendly fire involving Patriot systems, the missile system’s automated functions were identified as contributing factors in misidentifying friend as foe. U.S. Patriot batteries deployed to Iraq under the assumption that they would face heavy missile attacks, which would require the batteries to operate with a relative degree of autonomy in order to respond with sufficient speed. As a 2005 report by the Defense Science Board Task Force on the Patriot system’s performance observed, operating autonomously required U.S. forces to trust that the automated features of the system were functioning properly. So when the assumptions underlying the decision to allow the Patriot system to autonomously identify and sometimes fire on targets no longer applied, the soldiers operating the system were not in a position to question what the weapon’s sensors were telling them. Had U.S. and coalition forces faced heavy missile attacks in the war, automating such defenses would have made more sense. Instead, U.S. and allied forces quickly established air superiority, enough to drastically shift the balance of what was in the sky. Instead of facing large amounts of incoming missiles, Patriot batteries were observing large numbers of allied planes operating in the sky above them and sometimes struggling to identify friend from foe. According to the Defense Science Board’s task force, the first 30 days of combat in Iraq saw nine ballistic missile attacks that Patriot batteries might have been expected to counter, compared to 41,000 aircraft sorties, amounting to a “4,000-to-1 friendly-to-enemy ratio.” Picking out the correct targets against the background of a large number of potential false positives proved highly challenging. In the case of the Tornado shootdown, automation—and the speed with which automated action was taken—was likely sufficient on its own to cause the tragedy, but it might have been prevented if other systems hadn’t failed. As the UK Ministry of Defence concluded in its report examining the incident, the battery culpable for the shootdown was without its communications suite, which was still in transit from the United States. Contact with battalion headquarters occurred through a radio relay with another battery equipped with voice and data links to headquarters. “The lack of communications equipment meant that the Patriot crew did not have access to the widest possible ‘picture’ of the airspace around them to build situational awareness,” the report found. Another system that failed and that might have prevented the shootdown was the identification-as-friend-or-foe system, a safety measure designed to avoid such deadly mistakes. That kind of information, transmitted securely and immediately, could have prevented an automated system from shooting down the jet. If the information was communicated to the human crew operating the Patriot battery, it would have been a signal to call off the attack. Tragically, the IFF transponder or the Patriot battery’s ability to receive such a signal failed. While it is tempting to focus on the automated features of the Patriot system when examining the shootdown—or autonomous and semi-autonomous systems more broadly—it is important to consider such weapons as part of broader systems. As policymakers consider how to evaluate the deployment of increasingly autonomous weapons and military systems, the complexity of such systems, the ways in which they might fail, and how human operators oversee them are key issues to consider. Failures in communication, identification, and fire-control can occur at different points of a chain of events, and it can be difficult to predict how failures will interact with one another and produce a potentially lethal outcome. The Defense Science Board’s examination of the Patriot concluded that future conflicts will likely be “more stressing” and involve “simultaneous missile and air defense engagements.” In such a scenario, “a protocol that allows more operator oversight and control of major system actions will be needed,” the task force argued. Lessons learned since Finding the right mix of trust between an autonomous machine and the human relying on it is a delicate balance, especially given the inevitability of error. Seventeen years after the Tornado shootdown, the automated features of the Patriot missile remain in place, but the way in which they are used has shifted. Air threats, such as aircraft, helicopters, and cruise missiles can now only be engaged in manual mode “to reduce the risk of fratricide,” as the U.S. Army’s manual for air and missile defense outlines. In manual mode, automated systems still detect and track targets, but it’s a human who makes the call about when and if to fire. But “for ballistic missiles and anti-radiation missiles,” like the kind the Patriot in Iraq assumed the Tornado was, “the operator has a choice of engaging in the automatic or manual mode,” though the manual notes that these “engagements are typically conducted in the automatic mode.” Defense researchers caution that human beings are not well-suited to monitoring autonomous systems in this way. “Problems can arise when the automated control system has been developed because it presumably can do the job better than a human operator, but the operator is left in to ‘monitor’ that the automated system is performing correctly and intervene when it is not,” the engineering psychologist John Hawley, who was involved in the U.S. Army’s efforts to study the 2003 friendly fire incidents, wrote in a 2017 report. “Humans are very poor at meeting the monitoring and intervention demands imposed by supervisory control.” This dynamic played out in the other fatal friendly fire incident involving a Patriot missile battery during the Iraq War, when a U.S. Navy F/A-18 aircraft was misidentified as a ballistic missile and shot down, killing the pilot. According to a 2019 Center for Naval Analyses report, the Patriot recommended that the operator fire missiles in response to what it had identified as an enemy projectile, and the operator approved the recommendation to fire “without independent scrutiny of the information available to him.” This difficulty faced by Patriot missile batteries in correctly identifying potential targets illustrates one of the most serious challenges facing autonomous weapons—getting accurate training data. As militaries move toward greater autonomy in a wide range of systems, they are increasingly reliant on machine learning technology that uses large data sets to make predictions about how a machine should operate. The challenge of acquiring accurate data sets autonomous systems up for inevitable failure. “Conflict environments are harsh, dynamic and adversarial, and there will always be more variability in the real-world data of the battlefield than in the limited sample of data on which autonomous systems are built and verified,” as Arthur Holland Michel, and associate researcher in the Security and Technology Programme at the UN Institute for Disarmament Research, wrote in a report last year addressing data issues in military autonomous weapons. A lack of reliable data or an inability to produce datasets that replicate combat conditions will make it more likely that autonomous weapons fail to make accurate identifications. Aware of the potential for error, one way to adopt autonomous systems while addressing the risk to civilians and servicemembers is to shift toward a posture in which risk is borne primarily by the machine. The 2003 shootdowns involved Patriot missiles acting in self-defense and misidentifying their enemy. By accepting greater risk to autonomous systems—that they might be destroyed or disabled—autonomous systems can avoid the risk of friendly fire or civilian casualties by “using tactical patience, or allowing the platform to move in closer to get a more accurate determination of whether a threat actually exists,” as Larry Lewis, the author of the 2019 CNA report, argues. Rather than quickly firing in self-defense, this view argues for patience and sacrificing a measure of speed in favor of accuracy. More broadly, Lewis recommends a risk management approach to using AI. While the specific nature of every given error is hard to anticipate, the range of bad and undesired outcomes can fall in similar categories of error or outcome. Planning for AI incorporated into weapons, sensors, and information displays could include an awareness of error, and present that information in a useful way without adding to the cognitive load of the person using the machine. Artificial Intelligence has already moved beyond the speculative to tangible, real-world applications. It already informs the targeting decisions of military weapons, and will increasingly shape how people in combat use machines and tools. Adapting to this future, as the Pentagon and other military establishments seem intent to do, means planning for error, accidents, and novel harm, the way militaries have already adapted to such error in human hands. The Pentagon has taken some steps to address these risks. In February 2020, the Department of Defense released a set of principles AI ethics drafted by the Defense Innovation Board.

#### Military AI applications risk accuracy and information errors that threaten human dignity and civilian lives

Schmid et al ’22 (Schmid, S-- human-computer interaction cscw crisis informatics

., Riebe, T-- is a research associate and doctoral student at the Chair of Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science at the TU Darmstadt.

., & Reuter, C.-- Christian Reuter is Full Professor at Technical University of Darmstadt. His chair Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science with secondary appointment in the Department of History and Social Sciences combines computer science with peace and security research

(2022). “Dual-Use and Trustworthy? A Mixed Methods Analysis of AI Diffusion Between Civilian and Defense R&D.” Science and Engineering Ethics, 28(2). https://doi.org/10.1007/s11948-022-00364-7)//Marzz

Robustness, accuracy, and information quality seem to be apparent values which support Trustworthy AI, when considering military purposes. This does not mean that these norms are entirely absent when it comes to civilian AI applications. Instead, our analysis indicates that they are relatively more prevalent in the military context (see supplementary material Table A). Thus, as a value, robustness is comparatively more significant in the context of military (D11) applications, including resilience as an important standard (D10). Further, accuracy is particularly important in the context of military applications, including transparency on problems of inaccuracy: Although the RMS [root mean square; author’s note] errors for building reconstruction […] indicate that our method provides reasonable geometrical accuracies (height error is the same as for single points if the parallax accuracy is about one image pixel), the results in building detection are less precise. (D12.; own emp.) Similarly, the EU guideline stresses the importance of the technical values of robustness and accuracy. These relate to both safety and security, which are crucial in warfare scenarios. Military AI applications may support standards of Trustworthy AI, paying special attention to robustness and accuracy (European Commission, 2019) in more critical contexts. This reflects the potential to ensure security as proposed by the EU guideline (European Commission, 2019), while also indicating the technology’s possible normative ambiguity regarding general human and environmental well-being. Information quality has also been relatively more important for military (D1) applications. Given the high stake of a military operation, errors due to low information quality may have a greater impact on people, e.g., by mistaking civilian infrastructure for military bases or by falsely engaging civilians as combatants. Trustworthiness of Civilian AI Applications At the same time, there is a comparatively stronger interest in civilian projects in awareness, indicating the importance of capturing the environment in all its complexity. For example, SPARC, a project on autonomous driving in urban trafc, relies heavily on orientation in the context of moving and directing surrounding objects, opting for a “holistic representation” (D3), while at the same time training data is focused on “eventful […] and […] unique situations” (D13). Whether in terms of space, time, or speed, there is a strong reference to environmental information. This is surprising, as situational awareness is not only stressed by the EU (European Commission, 2019) but is mostly apparent in military contexts. Overall, civilian applications emphasize the relevance of explainability, which is referred to as “retaining many of the advantages of variational trajectory optimization methods, in particular expressiveness” (D11; own emphasis). Others underline that “[t]he ability for humans to understand the reasoning process is essential to the presented case study” (D13). This highlights the ambivalence of explainability as a normative concept. While it may be defned as the ability to explain, interpretability, namely the ability to provide (grounds for) an interpretation, is often associated with the concept of explainability, as it is also the case in the Trustworthy AI guide (European Commission, 2019). This requirement for civilian applications may be plausible, should special attention be paid to a broader and more diverse group of end-users. This becomes particularly apparent considering that the project on autonomous driving in cities (D11) stresses explainability (or expressiveness) the most. It should be noted that both security and safety were also qualitatively deduced regarding military applications, indicating human-centric approaches albeit in different terms. Human dignity, implying human-centric approaches, represents one of the core values of Trustworthy AI (European Commission, 2019). Such statements are more common in the context of civilian applications; as they apply AI applications that put focus on human reasoning, hand gestures, or the human body (see supplementary material Table A). Military applications accordingly refect less interest in a precise analysis of the social or intimate environment. Yet, a strong focus on people’s movements or behavior does not necessarily imply the implementation of a human-centric AI in terms of human dignity or personal rights. Difused Values Across Civilian and Military Applications Regardless of the feld of application, the authors of scientifc publications were transparent about procedural problems. In contrast, AI was depicted relatively fawless in online presentations of projects or product fyers. This may be due to the nature of scholarly debates, supporting values such as transparency (of problems). Problematic issues were not made transparent in shorter, more easily accessible online contributions, while such documents contained more direct references to economic merits. The European expert group’s guide would suggest presenting complex, inconvenient facts to a broader audience and allow for understandability independent from personal background (European Commission, 2019). Furthermore, the fgurative alignment of AI and animal behavior became visible. AI projects were oriented towards phenomena in nature, for example in the development of “swarms” of UAVs or processing as in an “ant colony” (D10). AI was also designed to imitate the human essence. This is refected in notions about the AI’s self and its abilities (see supplementary material Table A). Trustworthy AI refers to approaches such as values-by-design, implying a certain degree of technological agency (European Commission, 2019). However, Fraunhofer projects do not refect the awareness of such interactional approaches or non-human agency. While projects indicate anthropomorphization of AI as well as bionic models, they do not guarantee trustworthiness based on environmental awareness.

### Colonial

#### Data Colonialism Link – proliferation of AI technology thwarts the Global South into a new stage of coloniality, one in which Northern tech companies control economic and national sovereignty via a monopoly on data

Varon & Peña 21 (Joana, Executive Directress and Creative Chaos Catalyst at Coding Right, Paz Latin American Institute of Terraforming, “Artificial intelligence and consent: A feminist anticolonial critique”, Econstor, Institute for Internet and Society, Berlin, Vol. 10, Iss. 4, https://doi.org/10.14763/2021.4.1602 , pp. 15-17)-qcl

Indian digital anthropologist Payal Arora states (2019) that there is a tendency of states to experiment with people in economic vulnerability, as the damages that can be done are considered less important and it is more difficult for them to access justice for reparations. This extractivist logic focused on the most vulnerable prevails in AI systems developed for social welfare. They replicate what Couldry and Mejias (2018) call the "new state of capitalism" where the production and extraction of personal data naturalise the colonial appropriation of life in general. To achieve this, the authors consider that a series of ideological processes operate where, on the one hand, personal data is treated as raw material, naturally disposable for the expropriation of capital and, on the other, where corporations are considered the only ones capable of processing and, therefore, appropriate the data. Renata Ávila (2020) goes even further and points out how countries where most ‘big tech’ companies come from (the US and China, particularly) tend to benefit within a global system from the digitisation of poor and middle-income countries in what appears to be a new form of colonialism. Therefore, there is a multilayer of extractivism: at the level of individual countries, dominant elites processing, classifying and taking decisions about data of the poor, and at the global level, rich countries presenting themselves, and their companies, as the providers of “solutions”, benefiting from the profits of data colonialism. For this extractivist and data colonialist practices to prevail, there is a chain of subject focused consent from citizens to governments and from local governments to ‘big tech’. Once again, consent is being instrumentalised to enable for data processing even beyond data owner clear awareness of future usages and consequences. In the mentioned cases about Chile and Colombia, for example, there are private bidding processes. Moreover, in the case of the Colombian SISBEN, the state's bidding contract is part of a strategy to consolidate a data analysis market in Colombia, so that the Colombian company selected would provide a service to the state, while receiving training from MIT experts and access to a sufficiently massive database to experiment (López, 2020). In Latin America, IBM, Microsoft, NEC, Cisco, Google are commonly involved in AI projects developed by the public sector from the region. Every project feeds databases and provides intelligence for machine learning systems of these companies, which can use these less regulated environments, where enforcement of privacy rights is weak, as laboratories to test and improve their systems, normally unaccountable to possible harmful consequences. Who will own the knowledge and set the epistemologies of the categories running these AI systems? Very likely, the digital welfare hype in Latin America is feeding a circle in which a foreign agent, unaware of the context and with lived experience far different from the local culture will always be bringing what is commonly called “an innovative solution” to a problem, treated as something external and punctual, though most of those problems are historical, structural and actually caused or fed into by the actions of these very same corporations. With the inputs from these experiments, very commonly, these companies are also the vectors for spreading experiments from one country to another. For example, this is the case of Plataforma Tecnológica de Intervención Social (Technological Platform for Social Intervention), a machine learning experiment to predict teenage pregnancy and school dropouts, which was conducted by Microsoft, in partnership with the municipality of Salta, Argentina. “Intelligent algorithms allow us to identify characteristics in people that could end up with these problems and warn the government to work on their prevention,” said a Microsoft Azure representative in an interview for a company publication (News Center Microsoft Latinoamérica, 2018, n.p.). The system was heavily criticised due to statistical errors, sensitivities of reporting unwanted pregnancies, using data inadequate to make reliable predictions, but even further, for being used as a tool for discrimination of the poor and deviate the agenda of effective public policies to guarantee access to sexual and reproductive rights (Peña & Varon, 2019). Despite this, the programme is now being exported to other municipalities in Argentina, such as La Rioja, Tierra del Fuego, as well as to Colombia and Brazil (Peña & Varon, 2020).

#### AI impacts result on our society – decolonialization is a solution to challenge fairness and settled assumption

Muhammed 18 (Shakir, scientist and engineer in the fields of statistical machine learning and artificial intelligence, 10-11-2018, accessed on 6-24-2022, The Spectator, "Decolonising Artificial Intelligence", http://blog.shakirm.com/2018/10/decolonising-artificial-intelligence/)-qcl

AI will result in objects of culture, and its use will have impacts on the way our societies work (it already has). And this is why we should consider decolonisation seriously. But we are already engaged in these issues, under different headings and, importantly, in relation to their underlying technical questions. This brings the utility of decolonisation as a tool into question. The famous paper by Tuck and Yang entitled '[decolonisation is not a metaphor](https://www.researchgate.net/publication/277992187_Decolonization_Is_Not_a_Metaphor" \t "_blank)'[note]Tuck E, Yang KW. [Decolonization is not a metaphor](https://www.researchgate.net/publication/277992187_Decolonization_Is_Not_a_Metaphor). Decolonization: Indigeneity, education & society. 2012 Sep 8;1(1).[/note] enters: 'Decolonisation brings about repatriation of indigenous land and life; it is not a metaphor for other things we want to do to improve our societies and schools. E. Tuck and K. Wang, Decolonization is not a Metaphor, 2012. This is an important clarification. The key message they leave us with, is that it is problematic to use decolonisation as a placeholder for all the ways we wish to engage with social justice. It leads to a loss of meaning. Too often is decolonisation used for political symbolism. Too often are claims for decolonisation used to raise opposition, without genuine concerns. Too often is decolonisation used to signal an enemy. Our science will not be advanced through a world-view based on empty symbolism and opposition: together we can move beyond the too-easy narrative of them-vs-us, coloniser-vs-colonised, metropole-vs-south, west-vs-rest. More worryingly, decolonisation increasingly seems to be used as a replacement for Transformation. The [price of transformation](http://blog.shakirm.com/2018/09/the-price-of-transformation/) cannot be paid by allowing ourselves to be distracted by the language of decolonisation and delaying the work of deep social, institutional and personal responsibility for change. We can learn from the questions that decolonisation raises and the strategies it suggests. But we can also maintain our focus on the important questions of social justice. I will continue to demand for a simpler language. Let us say what we mean and want, to aim to be understood, to be more precise. And we can do this by keeping the challenges we identify and their scientific and technical basis in close proximity. There is a reason for concern! What else can we see when read in the [New York Times](https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html)[note]Kai Fu Lee, [The Real Threat of Artificial Intelligence](https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html), New York Times June 2017[/note] about one future path for our countries: "Unless they wish to plunge their people into poverty, they will be forced to negotiate with whichever country supplies most of their A.I. software — China or the United States — to essentially become that country’s economic dependent." Kai Fu Lee, [The Real Threat of Artificial Intelligence](https://www.nytimes.com/2017/06/24/opinion/sunday/artificial-intelligence-economic-inequality.html), June 2017 We immediately recognise the colonial nature of this possible future. When Ian Hogarth writes of [AI nationalism](https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism)[note]Ian Hogarth, [AI Nationalism](https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism), June 2018[/note], we recall the hubris and implications of the nations that sought Empire. We are not oblivious to the fact that a form of [imperialism based on data](https://modelviewculture.com/pieces/data-colonialism-critiquing-consent-and-control-in-tech-for-social-change)[note][Data Colonialism: Critiquing Consent and Control in “Tech for Social Change”](https://modelviewculture.com/pieces/data-colonialism-critiquing-consent-and-control-in-tech-for-social-change), Model View Culture, June 2016[/note] and its ownership is possible (if not underway). We do recognise a forming [cyber-colonialism](https://conspicuouschatter.wordpress.com/2014/06/21/the-dawn-of-cyber-colonialism/)[note][The Dawn of cyber-colonialism](https://conspicuouschatter.wordpress.com/2014/06/21/the-dawn-of-cyber-colonialism/), June 2014[/note] that expands as censorship increases and online freedoms are curtailed. This is where decolonisation plays its role. The solutions that have been tried in the restoration of land and life in the post-colonial age, can be ours to learn from and reuse. The basis of this solution is in self-ownership and its protection. Fortunately, as a field we have the basis of such protections already. We can continue to strengthen open-source software, open-data, and open-access science— publishing more, not less; we can further support accessible machine learning frameworks, and accessible scientific communication; and we can continue to find solutions to the challenges of fairness, privacy, safety, verification, and governance. And we can go further, by always challenging our settled assumptions and world-views as we expand the frontiers of our knowledge. The only AI that empowers and works for the benefit of humanity is a truly global AI. Making global AI truly global will not be easy. We have heard the call.

#### The transition to AI including further drives the Global South into economic dependency upon colonial powers

**Arun 19** (Chinmayi, Affiliate at Harvard at the Berkman Klein Center for Internet & Society & Fellow of the Information Society Project at Yale Law School, 07-20-19, “AI and the Global South: Designing for Other Worlds” Draft Chapter for the Oxford Handbook of Ethics of AI, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3403010, pp. 2-3)-qcl

There is an increasing awareness that we should be thinking more about the impact of AI on the Global South. The broad concern is clear enough: if privileged white men are designing the technology and the business models for AI 9 , how will they design for the South? The answer is that they will design in a manner that is at best an uneasy fit, and at worst amplifies existing systemic harm and oppression to horrifying proportions. As ‘Global South’ advocates furrow their brows about AI, they may be thinking of web-based AI designed by people who live in worlds that rarely see power cuts or internet shutdowns and then deployed to the rural hinterlands of countries with poor internet connectivity and only a few hours of electricity a day. They may worry about the resources diverted from education and health-care budgets to technology-centric solutions from the companies that are building these systems. They may be concerned about the surveillance of Southern children through AI for Education, built by people whose own children go to private school and have restricted access to screens. In authoritarian countries, they may lose sleep over AI that uses facial recognition, drones and other forms of surveillance to oppress vulnerable populations. They may worry about the loss of jobs and the impact on economies as AI replaces low-skilled workers. These concerns are not without foundation. Ideas of the past like one laptop per child10 have resulted in spectacular failure despite the bright-eyed optimism and laudable intentions with which they were created. Technology designed out of context may fail to take local resources, social norms and cultural context into account. 'One day delivery’ can mean very different things in Boston and Hyderabad even if the system designed for both cities is the same. Facebook can be fairly harmless in most countries and find itself weaponised in a country with Myanmar’s sociopolitical context, to contribute to genocide.11 It can take effort for Google Maps to be able to account for the favelas of Rio de Janeiro.12 Technology policy frameworks can impact whole countries, as we might have learned from the debate on drug patents and public health in developing world. There are so many ways in which Artificial Intelligence can wreak havoc in Southern countries and affect the human rights of Southern populations. In the absence of local regulation in Southern countries, AI may be deployed in its experimental stages such that the people of these countries bear the risk of harm that may ensue. At a larger scale, AI may impact the economies of these countries by affecting their role in the global economy: several developing countries that benefited from their role in the Internet-driven global economy may gradually find the low skilled outsourced services they offer replaced by automation. The ‘call-centres’ of Bangalore, and employment and business they generate, can be undone as automation makes human intervention unnecessary. Automated cars may result in the cab drivers of New York - famously from all over the world - finding themselves out of work with a redundant skill. We need to begin our journey towards including the South as a priority, and we need to do go beyond the mere use of the phrase in policy documents or speeches. For this, we have to understand the many things we specifically worry about when we speak of the Global South. Who is being left out and endangered?

#### The transition to AI including further drives the Global South into economic dependency upon colonial powers

**Arun 19** (Chinmayi, Affiliate at Harvard at the Berkman Klein Center for Internet & Society & Fellow of the Information Society Project at Yale Law School, 07-20-19, “AI and the Global South: Designing for Other Worlds” Draft Chapter for the Oxford Handbook of Ethics of AI, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3403010, pp. 2-3)-qcl

There is an increasing awareness that we should be thinking more about the impact of AI on the Global South. The broad concern is clear enough: if privileged white men are designing the technology and the business models for AI 9 , how will they design for the South? The answer is that they will design in a manner that is at best an uneasy fit, and at worst amplifies existing systemic harm and oppression to horrifying proportions. As ‘Global South’ advocates furrow their brows about AI, they may be thinking of web-based AI designed by people who live in worlds that rarely see power cuts or internet shutdowns and then deployed to the rural hinterlands of countries with poor internet connectivity and only a few hours of electricity a day. They may worry about the resources diverted from education and health-care budgets to technology-centric solutions from the companies that are building these systems. They may be concerned about the surveillance of Southern children through AI for Education, built by people whose own children go to private school and have restricted access to screens. In authoritarian countries, they may lose sleep over AI that uses facial recognition, drones and other forms of surveillance to oppress vulnerable populations. They may worry about the loss of jobs and the impact on economies as AI replaces low-skilled workers. These concerns are not without foundation. Ideas of the past like one laptop per child10 have resulted in spectacular failure despite the bright-eyed optimism and laudable intentions with which they were created. Technology designed out of context may fail to take local resources, social norms and cultural context into account. 'One day delivery’ can mean very different things in Boston and Hyderabad even if the system designed for both cities is the same. Facebook can be fairly harmless in most countries and find itself weaponised in a country with Myanmar’s sociopolitical context, to contribute to genocide.11 It can take effort for Google Maps to be able to account for the favelas of Rio de Janeiro.12 Technology policy frameworks can impact whole countries, as we might have learned from the debate on drug patents and public health in developing world. There are so many ways in which Artificial Intelligence can wreak havoc in Southern countries and affect the human rights of Southern populations. In the absence of local regulation in Southern countries, AI may be deployed in its experimental stages such that the people of these countries bear the risk of harm that may ensue. At a larger scale, AI may impact the economies of these countries by affecting their role in the global economy: several developing countries that benefited from their role in the Internet-driven global economy may gradually find the low skilled outsourced services they offer replaced by automation. The ‘call-centres’ of Bangalore, and employment and business they generate, can be undone as automation makes human intervention unnecessary. Automated cars may result in the cab drivers of New York - famously from all over the world - finding themselves out of work with a redundant skill. We need to begin our journey towards including the South as a priority, and we need to do go beyond the mere use of the phrase in policy documents or speeches. For this, we have to understand the many things we specifically worry about when we speak of the Global South. Who is being left out and endangered?

### Drones

#### **Drone warfare is replacing humans as sovereign decision makers—becomes an illusion where the god-trick is epistemological**

Wilcox ’16 (Lauren Wilcox, PhD, is a Staff Research Scientist in People & AI Research.--“Embodying algorithmic war: Gender, race, and the posthuman in drone warfare”—Article--Volume 48 Issue 1, February 2016--https://journals.sagepub.com/doi/abs/10.1177/0967010616657947)//Marzz

The use of weaponized drones to supplement and sometimes supplant ‘manned’ aircraft is at the forefront of debates over the use of algorithms, digital technologies and artificial intelligence in the projection of violence without the potential loss of human pilots. The algorithmic capabilities of data-driven technologies for the identification, localization, naming, and depiction of mobile targets have been theorized to enable certain geographies of security beyond the battlefield uses of algorithms and artificial intelligence (Amoore, 2009, 2013). However, even in the most direct and spectacular forms of violence associated with algorithms, such as their use in identifying and targeting individuals to assassinate via drone strikes, the question of the embodiment of decision-making remains vitally important. Discussions of artificial intelligence in war/ security practices have a tendency to focus on machines and technologies as ‘other than human’, caught in a zero-sum battle with humanity over the sovereign powers of life and death (Singer, 2009: 123–134; Berkowitz, 2014). While much of the debate over drone warfare is over the extent to which algorithms are replacing humans as sovereign decision-makers, the territorial expansion of the drone’s reach is also at issue. Donna Haraway (1988: 581) famously describes the ‘god-trick’ of Western scientific epistemologies: the illusion of being able to see everywhere from a disembodied position of ‘nowhere’ as an integral component of histories of militarism, capitalism, colonialism, and male supremacy. This ‘god-trick’ is seemingly perfected in the weaponized drone, with its global surveillance capacities and purported efficiency and accuracy in targeting weapons, and, as such, has been a frequent inspiration for critical work on the use of drones in warfare (Blanchard, 2011; Shaw and Akhtar, 2012; Stahl, 2013; Wilcox, 2015: 131–165). The ‘god-trick’ is not only visual, but more broadly epistemological: artificial intelligence, especially in an age of ‘big data’, can also appear to have omniscent power that appears everywhere and nowhere at once. Shaw (2012) warns, ‘Everywhere and nowhere, drones have become sovereign tools of life and death, and are coming to a sky near you’. Drone warfare, based on the algorithmic decision-making capacities of artificial intelligence and sophisticated visual surveillance, can seem to be an inhuman form of war in which bodies only appear as dead or dying victims, if they appear at all (Gregory, 2015). Grégoire Chamayou begins A Theory of the Drone (2014) by recounting the same massacre in the Afghan Uruzgan province that frames this piece, presenting a reading of the visual and computational powers of the drone as awesome and sublime: The eye of God, with its overhanging gaze, embraces the entire world. Its vision is more than just sight: beneath the skins of phenomena it can search hearts and minds. Nothing is opaque to it. Because it is eternity, it embraces the whole of time, the past as well as the future. (Chamayou, 2014: 37) Chamayou’s depiction of the drone as the ‘eye of God’ presents the death-dealing capacities of the drone as sovereign, able to see the entire world and into the past and future as well; creating archives of people’s lives and anticipating future movements (2014: 39–43). This vision of drone warfare has a long history in what Kaplan describes as the ‘cosmic view’ of airpower more broadly, in a ‘unifying gaze of an omniscient viewer of the global from a distance’ (Kaplan, 2006: 401), which plays a crucial role in the US imagination of its own national airspace as global but under threat. Haraway’s (1988) critique of the seeming transparency of visual technologies and their connection to epistemologies of domination rests upon the concept of embodiment of all vision and calls for a feminist project of partiality, structuring, and situating. Such a project has important resonances with Walters’s call for a ‘zonal’ rather than ‘global’ theorization of emergent spaces of security attuned to the uneven particularities and multiplicities of bordering practices, rather than the smooth homogeneities of the ‘global’ (2011). As an extension of airpower, drone warfare is a practice of bordering that takes place in a frontier logic of the extension of American sovereignty ‘vertically’ though the air, but with an ambiguous relationship to imperialism on the ground (with drones operating over more states than Americans have troops on the ground, as in Afghanistan). ‘With whose blood were my eyes crafted?’ (Haraway, 1988: 585). In this statement, Haraway provocatively asserts that visual practices are situated, embodied, and both enabled by and enable violent practices of domination. In the embodying assemblages of drone warfare, as a form of necropolitics, ‘it makes little sense to insist on distinctions between “internal” and “external” political realms, separated by clearly demarcated boundaries’ (Mbembe, 2003: 32). In this piece I argue that ‘drone assemblages’ as a mode of necropolitical violence – the violence of ‘distinguishing whose life is to be managed and those who are subject to the right of death’ (Allinson, 2015: 121; Mbembe, 2003) – is both a form of posthuman embodiment and is itself corporealizing in terms of the racialized and gendered bodies it produces as either killable or manageable. As such, an embodied reading of drone warfare suggests the limits of the ‘god-trick’ of drone warfare both in terms of its omniscient surveillance capacities as well as its global spread. An embodied reading further contributes to critical analyses of targeting practices, such as that of Zehfuss (2011), that undermine claims of the precision and discrimination of such technological practices.

### Civilian AI

#### Trustworthy AI values relate to safety and security and reflects of human essence ---although they are not trustworthy based on environmental awareness

Schmid et al ’22 (Schmid, S-- human-computer interaction cscw crisis informatics

., Riebe, T-- is a research associate and doctoral student at the Chair of Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science at the TU Darmstadt.

., & Reuter, C.-- Christian Reuter is Full Professor at Technical University of Darmstadt. His chair Science and Technology for Peace and Security (PEASEC) in the Department of Computer Science with secondary appointment in the Department of History and Social Sciences combines computer science with peace and security research

(2022). “Dual-Use and Trustworthy? A Mixed Methods Analysis of AI Diffusion Between Civilian and Defense R&D.” Science and Engineering Ethics, 28(2). https://doi.org/10.1007/s11948-022-00364-7)//Marzz

Robustness, accuracy, and information quality seem to be apparent values which support Trustworthy AI, when considering military purposes. This does not mean that these norms are entirely absent when it comes to civilian AI applications. Instead, our analysis indicates that they are relatively more prevalent in the military context (see supplementary material Table A). Thus, as a value, robustness is comparatively more significant in the context of military (D11) applications, including resilience as an important standard (D10). Further, accuracy is particularly important in the context of military applications, including transparency on problems of inaccuracy: Although the RMS [root mean square; author’s note] errors for building reconstruction […] indicate that our method provides reasonable geometrical accuracies (height error is the same as for single points if the parallax accuracy is about one image pixel), the results in building detection are less precise. (D12.; own emp.) Similarly, the EU guideline stresses the importance of the technical values of robustness and accuracy. These relate to both safety and security, which are crucial in warfare scenarios. Military AI applications may support standards of Trustworthy AI, paying special attention to robustness and accuracy (European Commission, 2019) in more critical contexts. This reflects the potential to ensure security as proposed by the EU guideline (European Commission, 2019), while also indicating the technology’s possible normative ambiguity regarding general human and environmental well-being. Information quality has also been relatively more important for military (D1) applications. Given the high stake of a military operation, errors due to low information quality may have a greater impact on people, e.g., by mistaking civilian infrastructure for military bases or by falsely engaging civilians as combatants. Trustworthiness of Civilian AI Applications At the same time, there is a comparatively stronger interest in civilian projects in awareness, indicating the importance of capturing the environment in all its complexity. For example, SPARC, a project on autonomous driving in urban trafc, relies heavily on orientation in the context of moving and directing surrounding objects, opting for a “holistic representation” (D3), while at the same time training data is focused on “eventful […] and […] unique situations” (D13). Whether in terms of space, time, or speed, there is a strong reference to environmental information. This is surprising, as situational awareness is not only stressed by the EU (European Commission, 2019) but is mostly apparent in military contexts. Overall, civilian applications emphasize the relevance of explainability, which is referred to as “retaining many of the advantages of variational trajectory optimization methods, in particular expressiveness” (D11; own emphasis). Others underline that “[t]he ability for humans to understand the reasoning process is essential to the presented case study” (D13). This highlights the ambivalence of explainability as a normative concept. While it may be defined as the ability to explain, interpretability, namely the ability to provide (grounds for) an interpretation, is often associated with the concept of explainability, as it is also the case in the Trustworthy AI guide (European Commission, 2019). This requirement for civilian applications may be plausible, should special attention be paid to a broader and more diverse group of end-users. This becomes particularly apparent considering that the project on autonomous driving in cities (D11) stresses explainability (or expressiveness) the most. It should be noted that both security and safety were also qualitatively deduced regarding military applications, indicating human-centric approaches albeit in different terms. Human dignity, implying human-centric approaches, represents one of the core values of Trustworthy AI (European Commission, 2019). Such statements are more common in the context of civilian applications; as they apply AI applications that put focus on human reasoning, hand gestures, or the human body (see supplementary material Table A). Military applications accordingly refect less interest in a precise analysis of the social or intimate environment. Yet, a strong focus on people’s movements or behavior does not necessarily imply the implementation of a human-centric AI in terms of human dignity or personal rights. Difused Values Across Civilian and Military Applications Regardless of the feld of application, the authors of scientifc publications were transparent about procedural problems. In contrast, AI was depicted relatively fawless in online presentations of projects or product fyers. This may be due to the nature of scholarly debates, supporting values such as transparency (of problems). Problematic issues were not made transparent in shorter, more easily accessible online contributions, while such documents contained more direct references to economic merits. The European expert group’s guide would suggest presenting complex, inconvenient facts to a broader audience and allow for understandability independent from personal background (European Commission, 2019). Furthermore, the fgurative alignment of AI and animal behavior became visible. AI projects were oriented towards phenomena in nature, for example in the development of “swarms” of UAVs or processing as in an “ant colony” (D10). AI was also designed to imitate the human essence. This is refected in notions about the AI’s self and its abilities (see supplementary material Table A). Trustworthy AI refers to approaches such as values-by-design, implying a certain degree of technological agency (European Commission, 2019). However, Fraunhofer projects do not refect the awareness of such interactional approaches or non-human agency. While projects indicate anthropomorphization of AI as well as bionic models, they do not guarantee trustworthiness based on environmental awareness.

### AI racist

#### AI inevitably reinforces systems of discrimination – using datasets from discriminatory practices and differentiating individuals based on identities it leads to discrimination especially for governments in the Global South who can’t fight back against Northern tech companies

**Arun 19** (Chinmayi, Affiliate at Harvard at the Berkman Klein Center for Internet and Society & Fellow of the Information Society Project at Yale Law School, 07-20-19, “AI and the Global South: Designing for Other Worlds” Draft Chapter for the Oxford Handbook of Ethics of AI, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3403010, pp. 9-12)-qcl

It is worth reading work by scholars who think about AI and discrimination, while noting that Southern institutions and legal frameworks can exacerbate the harms that they discuss. Southern populations within Northern countries might not have same access as privileged people to the institutions within the same countries. Autonomous systems are used so broadly that they can affect the economy, housing, intimate relationships and more. They can introduce or enhance discrimination and oppression, and they can erase populations by failing to account for their existence. I begin with discussing autonomous systems as systems of discrimination, I then move on to discussing what this may mean for Southern populations, especially since the fragile democracies and non-democracies of the world do not offer their citizens the institutional protections that may be available in the USA or Europe. Any discussion of AI in the context of discrimination has to discuss big data, which is ‘the fuel that runs the Algorithmic Society’. Algorithmic systems are often trained on a corpus of data, which means that the big data and its inherent biases affect the outcome of these systems.61 There are several stages at which inaccuracies and bias can be introduced into algorithmic decisionmaking. These range from the recording of the data to the actual question answered by the algorithm. There is a tendency to accept predictions based on datasets as the truth62 even though the the outcome is typically an interpretation of the data63 and may be inaccurate.64 The dataset could suffer from any number of problems which would skew the outcome. Scholars use the term ‘dirty data’ to refer to missing, incorrect and badly represented data, as well as to data that has been manipulated intentionally or distorted by biases.65 Crawford has pointed out that “not all data is created or even collected equally”.66 Data collection has embedded power and assumptions. The recording of fingerprints for example is difficult for those who do manual work such as refugees, and migrant and contract labourers.67 The very design of data sets can be biased as a result of assumptions and gaps.68 The datasets could under-represent or wrongly represent certain populations, leading to discrimination against them or to their exclusion.69 Even if the dataset is accurate, its structure can end up discriminating and marginalising people: the classic example being datasets that code people as either male or female, erasing other forms of gender identity.70 A dataset might discriminate indirectly by recording a seemingly innocuous fact that acts as a marker for identity. An illustration of this is employment which can be used to infer caste based on the historic employment of marginalised caste people for certain tasks (such as manual scavenging).71 The training data for algorithms can embed bias,72and algorithms trained on real world data would replicate real word discrimination.73 Therefore a hospital computer program used to sort out medical school students based on previous admissions decisions, ends up discriminating against women and racial minorities because of the rules it learned from the hospital’s older biased decisions. 74 Big data essentially generates correlations. 75 Although scientists understand the difference between correlation and causation, the rest of the world tends to treat conclusions based on big data as ‘enough’.76 The AI Now institute has articulated the problem in unambiguous terms.77 It has pointed out that since classification, differentiation and ranking are central to AI systems, these systems are ‘systems of discrimination’. It has argued that the bias in AI systems is connected with the lack of diversity in the AI industry, including the people who build AI tools and the environment in which they are built. The large scale AI systems come from elite university labs and a few technology companies, which are ‘white, affluent, technically oriented and male’ spaces.78 In other words, these technologies are designed by people from the North. Context can be reintroduced if universities studying AI collaborate with social and humanities disciplines, affected communities and civil society organisations. 79 It is important in to account for plurality, context and intersectionality.80 In addition to changing how decisions are made about design, data and deployment in the algorithmic society, we must give Southern populations the tools to engage effectively with the questions that affect them. This is already proving challenging in what we understand as Global North countries despite the lively debate and relatively strong privacy and anti-discrimination laws. When companies deploy these technologies in Southern countries, there are fewer resources and institutions to help protect marginalised people’s rights. This needs to be remedied as a high priority The systems discussed in the four case studies are designed by people with privileged access to the data of data subjects. The data subjects have little control or autonomy over their own data. It is typical, when autonomous systems are used, that the data subject has no idea who has access to their data or how it is used. 81 This is exacerbated in Southern countries. Young democracies lack institutional stability since it takes time to build institutions and institutionalise democratic practices.82 This is why Milan argues that we need diverse ways for citizens and civil society engagement to ward off datafication practices that result in oppression and inequality.83 The institutional frameworks of Southern countries must be taken into account as we consider what impact AI might have on the South. Freedom depends not just on political and civil rights, but also on other social and economic arrangement such as education and health care. 84 Development, Amartya Sen argues, depends on the removal of sources ‘unfreedom’ such as systematic social deprivation, poverty, poor economic opportunities and tyranny, Sen describes poverty in terms of capability deprivation, in what is now famously knows as the ‘capabilities approach’ to development. Julie Cohen has applied Sen’s work, as build on by Martha Nussbaum, to access to knowledge, and has pointed out that we need to pay more attention to the relationship between the networked information environment and human flourishing.85 The rights of Southern populations can be realised through efforts made by states, but can also be eroded by the governing elite of states. In the past, Southern countries worked together as a bloc, to gain access to technology, capital and markets. 86 They had a shared commitment to development, opposition of colonialism, the creation of equitable conditions for socio-economic development of all countries and the evolution of South-South co operation. 87This co-operation has been taking place since the Non Aligned movement, in which developing countries came together to negotiate development and trade issues. As the developing countries began what they called South-South co-operation, triangular co-operation also began such that donors and northern partners became involved in South-South initiatives.88 Progress has been made over the years on South-South initiatives but one might argue that the cooperation between the Southern states and triangular co-operation has had mixed results. Over the years, non State actors such as businesses and civil society have started playing a powerful role in Southern Countries. These countries have developed groups that are wealthy and influential, and populations that are more affluent than their fellow citizens - the extractive, exploitative consequences are evident in the Aadhaar case study. Some Southern states are more developed and have greater economic influence than other Southern states. The exploitative nature of this relationship is evident in the China-Zimbabwe case study.

#### AI inevitably reinforces systems of discrimination – using datasets from discriminatory practices and differentiating individuals based on identities it leads to discrimination especially for governments in the Global South who can’t fight back against Northern tech companies

**Arun 19** (Chinmayi, Affiliate at Harvard at the Berkman Klein Center for Internet and Society & Fellow of the Information Society Project at Yale Law School, 07-20-19, “AI and the Global South: Designing for Other Worlds” Draft Chapter for the Oxford Handbook of Ethics of AI, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3403010, pp. 9-12)-qcl

It is worth reading work by scholars who think about AI and discrimination, while noting that Southern institutions and legal frameworks can exacerbate the harms that they discuss. Southern populations within Northern countries might not have same access as privileged people to the institutions within the same countries. Autonomous systems are used so broadly that they can affect the economy, housing, intimate relationships and more. They can introduce or enhance discrimination and oppression, and they can erase populations by failing to account for their existence. I begin with discussing autonomous systems as systems of discrimination, I then move on to discussing what this may mean for Southern populations, especially since the fragile democracies and non-democracies of the world do not offer their citizens the institutional protections that may be available in the USA or Europe. Any discussion of AI in the context of discrimination has to discuss big data, which is ‘the fuel that runs the Algorithmic Society’. Algorithmic systems are often trained on a corpus of data, which means that the big data and its inherent biases affect the outcome of these systems.61 There are several stages at which inaccuracies and bias can be introduced into algorithmic decisionmaking. These range from the recording of the data to the actual question answered by the algorithm. There is a tendency to accept predictions based on datasets as the truth62 even though the the outcome is typically an interpretation of the data63 and may be inaccurate.64 The dataset could suffer from any number of problems which would skew the outcome. Scholars use the term ‘dirty data’ to refer to missing, incorrect and badly represented data, as well as to data that has been manipulated intentionally or distorted by biases.65 Crawford has pointed out that “not all data is created or even collected equally”.66 Data collection has embedded power and assumptions. The recording of fingerprints for example is difficult for those who do manual work such as refugees, and migrant and contract labourers.67 The very design of data sets can be biased as a result of assumptions and gaps.68 The datasets could under-represent or wrongly represent certain populations, leading to discrimination against them or to their exclusion.69 Even if the dataset is accurate, its structure can end up discriminating and marginalising people: the classic example being datasets that code people as either male or female, erasing other forms of gender identity.70 A dataset might discriminate indirectly by recording a seemingly innocuous fact that acts as a marker for identity. An illustration of this is employment which can be used to infer caste based on the historic employment of marginalised caste people for certain tasks (such as manual scavenging).71 The training data for algorithms can embed bias,72and algorithms trained on real world data would replicate real word discrimination.73 Therefore a hospital computer program used to sort out medical school students based on previous admissions decisions, ends up discriminating against women and racial minorities because of the rules it learned from the hospital’s older biased decisions. 74 Big data essentially generates correlations. 75 Although scientists understand the difference between correlation and causation, the rest of the world tends to treat conclusions based on big data as ‘enough’.76 The AI Now institute has articulated the problem in unambiguous terms.77 It has pointed out that since classification, differentiation and ranking are central to AI systems, these systems are ‘systems of discrimination’. It has argued that the bias in AI systems is connected with the lack of diversity in the AI industry, including the people who build AI tools and the environment in which they are built. The large scale AI systems come from elite university labs and a few technology companies, which are ‘white, affluent, technically oriented and male’ spaces.78 In other words, these technologies are designed by people from the North. Context can be reintroduced if universities studying AI collaborate with social and humanities disciplines, affected communities and civil society organisations. 79 It is important in to account for plurality, context and intersectionality.80 In addition to changing how decisions are made about design, data and deployment in the algorithmic society, we must give Southern populations the tools to engage effectively with the questions that affect them. This is already proving challenging in what we understand as Global North countries despite the lively debate and relatively strong privacy and anti-discrimination laws. When companies deploy these technologies in Southern countries, there are fewer resources and institutions to help protect marginalised people’s rights. This needs to be remedied as a high priority The systems discussed in the four case studies are designed by people with privileged access to the data of data subjects. The data subjects have little control or autonomy over their own data. It is typical, when autonomous systems are used, that the data subject has no idea who has access to their data or how it is used. 81 This is exacerbated in Southern countries. Young democracies lack institutional stability since it takes time to build institutions and institutionalise democratic practices.82 This is why Milan argues that we need diverse ways for citizens and civil society engagement to ward off datafication practices that result in oppression and inequality.83 The institutional frameworks of Southern countries must be taken into account as we consider what impact AI might have on the South. Freedom depends not just on political and civil rights, but also on other social and economic arrangement such as education and health care. 84 Development, Amartya Sen argues, depends on the removal of sources ‘unfreedom’ such as systematic social deprivation, poverty, poor economic opportunities and tyranny, Sen describes poverty in terms of capability deprivation, in what is now famously knows as the ‘capabilities approach’ to development. Julie Cohen has applied Sen’s work, as build on by Martha Nussbaum, to access to knowledge, and has pointed out that we need to pay more attention to the relationship between the networked information environment and human flourishing.85 The rights of Southern populations can be realised through efforts made by states, but can also be eroded by the governing elite of states. In the past, Southern countries worked together as a bloc, to gain access to technology, capital and markets. 86 They had a shared commitment to development, opposition of colonialism, the creation of equitable conditions for socio-economic development of all countries and the evolution of South-South co operation. 87This co-operation has been taking place since the Non Aligned movement, in which developing countries came together to negotiate development and trade issues. As the developing countries began what they called South-South co-operation, triangular co-operation also began such that donors and northern partners became involved in South-South initiatives.88 Progress has been made over the years on South-South initiatives but one might argue that the cooperation between the Southern states and triangular co-operation has had mixed results. Over the years, non State actors such as businesses and civil society have started playing a powerful role in Southern Countries. These countries have developed groups that are wealthy and influential, and populations that are more affluent than their fellow citizens - the extractive, exploitative consequences are evident in the Aadhaar case study. Some Southern states are more developed and have greater economic influence than other Southern states. The exploitative nature of this relationship is evident in the China-Zimbabwe case study.

#### AI is proposed as the solution to implicit bias but in reality it reproduces systems of oppression through its very nature

**Buranyi 17** (Stephen, writer specialising in science and the environment, 8-8-2017, accessed on 6-21-2022, the Guardian, "Rise of the racist robots – how AI is learning all our worst impulses", <https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses)-qcl>

But, while some of the most prominent voices in the industry are concerned with the far-off future apocalyptic potential of AI, there is less attention paid to the more immediate problem of how we prevent these programs from amplifying the inequalities of our past and affecting the most vulnerable members of our society. When the data we feed the machines reflects the history of our own unequal society, we are, in effect, asking the program to learn our own biases. “If you’re not careful, you risk automating the exact same biases these programs are supposed to eliminate,” says Kristian Lum, the lead statistician at the San Francisco-based, non-profit Human Rights Data Analysis Group (HRDAG). Last year, Lum and a co-author showed that PredPol, a program for police departments that predicts hotspots where future crime might occur, could potentially get stuck in a feedback loop of over-policing majority black and brown neighbourhoods. The program was “learning” from previous crime reports. For Samuel Sinyangwe, a justice activist and policy researcher, this kind of approach is “especially nefarious” because police can say: “We’re not being biased, we’re just doing what the math tells us.” And the public perception might be that the algorithms are impartial. We have already seen glimpses of what might be on the horizon. Programs developed by companies at the forefront of AI research have resulted in a string of errors that look uncannily like the darker biases of humanity: a Google image recognition program labelled the faces of several black people as gorillas; a LinkedIn advertising program showed a preference for male names in searches, and a Microsoft chatbot called Tay spent a day learning from Twitter and began spouting antisemitic messages. These small-scale incidents were all quickly fixed by the companies involved and have generally been written off as “gaffes”. But the Compas revelation and Lum’s study hint at a much bigger problem, demonstrating how programs could replicate the sort of large-scale systemic biases that people have spent decades campaigning to educate or legislate away. Computers don’t become biased on their own. They need to learn that from us. For years, the vanguard of computer science has been working on machine learning, often having programs learn in a similar way to humans – observing the world (or at least the world we show them) and identifying patterns. In 2012, Google researchers fed their computer “brain” millions of images from YouTube videos to see what it could recognise. It responded with blurry black-and-white outlines of human and cat faces. The program was never given a definition of a human face or a cat; it had observed and “learned” two of our favourite subjects. This sort of approach has allowed computers to perform tasks – such as language translation, recognising faces or recommending films in your Netflix queue – that just a decade ago would have been considered too complex to automate. But as the algorithms learn and adapt from their original coding, they become more opaque and less predictable. It can soon become difficult to understand exactly how the complex interaction of algorithms generated a problematic result. And, even if we could, private companies are disinclined to reveal the commercially sensitive inner workings of their algorithms (as was the case with Northpointe). Less difficult is predicting where problems can arise. Take Google’s face recognition program: cats are uncontroversial, but what if it was to learn what British and American people think a CEO looks like? The results would likely resemble the near-identical portraits of older white men that line any bank or corporate lobby. And the program wouldn’t be inaccurate: only 7% of FTSE CEOs are women. Even fewer, just 3%, have a BME background. When computers learn from us, they can learn our less appealing attributes. Joanna Bryson, a researcher at the University of Bath, studied a program designed to “learn” relationships between words. It trained on millions of pages of text from the internet and began clustering female names and pronouns with jobs such as “receptionist” and “nurse”. Bryson says she was astonished by how closely the results mirrored the real-world gender breakdown of those jobs in US government data, a nearly 90% correlation. “People expected AI to be unbiased; that’s just wrong. If the underlying data reflects stereotypes, or if you train AI from human culture, you will find these things,” Bryson says. So who stands to lose out the most? Cathy O’Neil, the author of the book Weapons of Math Destruction about the dangerous consequences of outsourcing decisions to computers, says it’s generally the most vulnerable in society who are exposed to evaluation by automated systems. A rich person is unlikely to have their job application screened by a computer, or their loan request evaluated by anyone other than a bank executive. In the justice system, the thousands of defendants with no money for a lawyer or other counsel would be the most likely candidates for automated evaluation. In London, Hackney council has recently been working with a private company to apply AI to data, including government health and debt records, to help predict which families have children at risk of ending up in statutory care. Other councils have reportedly looked into similar programs. In her 2016 paper, HRDAG’s Kristian Lum demonstrated who would be affected if a program designed to increase the efficiency of policing was let loose on biased data. Lum and her co-author took PredPol – the program that suggests the likely location of future crimes based on recent crime and arrest statistics – and fed it historical drug-crime data from the city of Oakland’s police department. PredPol showed a daily map of likely “crime hotspots” that police could deploy to, based on information about where police had previously made arrests. The program was suggesting majority black neighbourhoods at about twice the rate of white ones, despite the fact that when the statisticians modelled the city’s likely overall drug use, based on national statistics, it was much more evenly distributed. As if that wasn’t bad enough, the researchers also simulated what would happen if police had acted directly on PredPol’s hotspots every day and increased their arrests accordingly: the program entered a feedback loop, predicting more and more crime in the neighbourhoods that police visited most. That caused still more police to be sent in. It was a virtual mirror of the real-world criticisms of initiatives such as New York City’s controversial “stop-and-frisk” policy. By over-targeting residents with a particular characteristic, police arrested them at an inflated rate, which then justified further policing. PredPol’s co-developer, Prof Jeff Brantingham, acknowledged the concerns when asked by the Washington Post. He claimed that – to combat bias – drug arrests and other offences that rely on the discretion of officers were not used with the software because they are often more heavily enforced in poor and minority communities. And while most of us don’t understand the complex code within programs such as PredPol, Hamid Khan, an organiser with Stop LAPD Spying Coalition, a community group addressing police surveillance in Los Angeles, says that people do recognise predictive policing as “another top-down approach where policing remains the same: pathologising whole communities”. There is a saying in computer science, something close to an informal law: garbage in, garbage out. It means that programs are not magic. If you give them flawed information, they won’t fix the flaws, they just process the information. Khan has his own truism: “It’s racism in, racism out.” It’s unclear how existing laws to protect against discrimination and to regulate algorithmic decision-making apply in this new landscape. Often the technology moves faster than governments can address its effects. In 2016, the Cornell University professor and former Microsoft researcher Solon Barocas claimed that current laws “largely fail to address discrimination” when it comes to big data and machine learning. Barocas says that many traditional players in civil rights, including the American Civil Liberties Union (ACLU), are taking the issue on in areas such as housing or hiring practices. Sinyangwe recently worked with the ACLU to try to pass city-level policies requiring police to disclose any technology they adopt, including AI.

#### AI contains human biases that are disproportionately used against minority communities

Pazzanese 20 Christina Pazzanese, writer for the Harvard Gazette, 10-26-2020, "Ethical concerns mount as AI takes bigger decision-making role," Harvard Gazette, <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/> [AJL]

A VENEER OF OBJECTIVITY Not everyone sees blue skies on the horizon, however. Many worry whether the coming age of AI will bring new, faster, and frictionless ways to discriminate and divide at scale. “Part of the appeal of algorithmic decision-making is that it seems to offer an objective way of overcoming human subjectivity, bias, and prejudice,” said political philosopher Michael Sandel, Anne T. and Robert M. Bass Professor of Government. “But we are discovering that many of the algorithms that decide who should get parole, for example, or who should be presented with employment opportunities or housing … replicate and embed the biases that already exist in our society.” Karen Mills. Jon Chase/Harvard file photo “If we’re not thoughtful and careful, we’re going to end up with redlining again.” — Karen Mills, senior fellow at the Business School and head of the U.S. Small Business Administration from 2009 to 2013 AI presents three major areas of ethical concern for society: privacy and surveillance, bias and discrimination, and perhaps the deepest, most difficult philosophical question of the era, the role of human judgment, said Sandel, who teaches a course in the moral, social, and political implications of new technologies. “Debates about privacy safeguards and about how to overcome bias in algorithmic decision-making in sentencing, parole, and employment practices are by now familiar,” said Sandel, referring to conscious and unconscious prejudices of program developers and those built into datasets used to train the software. “But we’ve not yet wrapped our minds around the hardest question: Can smart machines outthink us, or are certain elements of human judgment indispensable in deciding some of the most important things in life?” Panic over AI suddenly injecting bias into everyday life en masse is overstated, says Fuller. First, the business world and the workplace, rife with human decision-making, have always been riddled with “all sorts” of biases that prevent people from making deals or landing contracts and jobs. When calibrated carefully and deployed thoughtfully, resume-screening software allows a wider pool of applicants to be considered than could be done otherwise, and should minimize the potential for favoritism that comes with human gatekeepers, Fuller said. Sandel disagrees. “AI not only replicates human biases, it confers on these biases a kind of scientific credibility. It makes it seem that these predictions and judgments have an objective status,” he said. In the world of lending, algorithm-driven decisions do have a potential “dark side,” Mills said. As machines learn from data sets they’re fed, chances are “pretty high” they may replicate many of the banking industry’s past failings that resulted in systematic disparate treatment of African Americans and other marginalized consumers. “If we’re not thoughtful and careful, we’re going to end up with redlining again,” she said. A highly regulated industry, banks are legally on the hook if the algorithms they use to evaluate loan applications end up inappropriately discriminating against classes of consumers, so those “at the top levels” in the field are “very focused” right now on this issue, said Mills, who closely studies the rapid changes in financial technology, or “fintech.” “They really don’t want to discriminate. They want to get access to capital to the most creditworthy borrowers,” she said. “That’s good business for them, too.” OVERSIGHT OVERWHELMED Given its power and expected ubiquity, some argue that the use of AI should be tightly regulated. But there’s little consensus on how that should be done and who should make the rules. Thus far, companies that develop or use AI systems largely self-police, relying on existing laws and market forces, like negative reactions from consumers and shareholders or the demands of highly-prized AI technical talent to keep them in line. “There’s no businessperson on the planet at an enterprise of any size that isn’t concerned about this and trying to reflect on what’s going to be politically, legally, regulatorily, [or] ethically acceptable,” said Fuller. Firms already consider their own potential liability from misuse before a product launch, but it’s not realistic to expect companies to anticipate and prevent every possible unintended consequence of their product, he said. Few think the federal government is up to the job, or will ever be. “The regulatory bodies are not equipped with the expertise in artificial intelligence to engage in [oversight] without some real focus and investment,” said Fuller, noting the rapid rate of technological change means even the most informed legislators can’t keep pace. Requiring every new product using AI to be prescreened for potential social harms is not only impractical, but would create a huge drag on innovation.

#### AI used to exacerbate racism

Schneier 16 Bruce Schneier, 02-08-2016, " Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World" W. W. Norton & Company, [https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath\_-2015.pdf pg 780](https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath_-2015.pdf%20pg%20-77-780)[AJL]

In a fundamental way, companies use surveillance data to discriminate. They place people into different categories and market goods and services to them differently on the basis of those categories. “Redlining” is a term from the 1960s to describe a practice that’s much older: banks discriminating against members of minority groups when they tried to purchase homes. Banks would not approve mortgages in minority neighborhoods—they would draw a red line on their maps delineating those zones. Or they would issue mortgages to minorities only if they were buying houses in predominantly minority neighborhoods. It’s illegal, of course, but for a long time banks got away with it. More generally, redlining is the practice of denying or charging more for services by using neighborhood as a proxy for race—and it’s much easier to do on the Internet. In 2000, Wells Fargo bank created a website to promote its home mortgages. The site featured a “community calculator” to help potential buyers search for neighborhoods. The calculator collected the current ZIP code of the potential customers and steered them to neighborhoods based on the predominant race of that ZIP code. The site referred white residents to white neighborhoods, and black residents to black neighborhoods. This practice is called weblining, and it has the potential to be much more pervasive and much more discriminatory than traditional redlining. Because corporations collect so much data about us and can compile such detailed profiles, they can influence us in many different ways. A 2014 White House report on big data concluded, “… big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.” I think the report understated the risk. Price discrimination is also a big deal these days. It’s not discrimination in the same classic racial or gender sense as weblining; it’s companies charging different people different prices to realize as much profit as possible. We’re most familiar with this concept with respect to airline tickets. Prices change all the time, and depend on factors like how far in advance we purchase, what days we’re traveling, and how full the flight is. The airline’s goal is to sell tickets to vacationers at the bargain prices they’re willing to pay, while at the same time extracting from business travelers the much higher amounts that they’re willing to pay. There is nothing nefarious about the practice; it’s just a way of maximizing revenues and profits. Even so, price discrimination can be very unpopular. Raising the price of snow shovels after a snowstorm, for example, is considered pricegouging. This is why it is often cloaked in things like special offers, coupons, or rebates. Some types of price discrimination are illegal. For example, a restaurant cannot charge different prices depending on the gender or race of the customer. But it can charge different prices based on time of day, which is why you see lunch and dinner menus with the same items and different prices. Offering senior discounts and special children’s menus is legal price discrimination. Uber’s surge pricing is also legal. In many industries, the options you’re offered, the price you pay, and the service you receive depend on information about you: bank loans, auto insurance, credit cards, and so on. Internet surveillance facilitates a fine-tuning of this practice. Online merchants already show you different prices and options based on your history and what they know about you. Depending on who you are, you might see a picture of a red convertible or a picture of a minivan in online car ads, and be offered different options for financing and discounting when you visit dealer websites. According to a 2010 Wall Street Journal article, the price you pay on the Staples website depends on where you are located, and how close a competitor’s store is to you. The article states that other companies, like Rosetta Stone and Home Depot, are also adjusting prices on the basis of information about the individual user. More broadly, we all have a customer score. Data brokers assign it to us. It’s like a credit score, but it’s not a single number, and it’s focused on what you buy, based on things like purchasing data from retailers, personal financial information, survey data, warranty card registrations, social media interactions, loyalty card data, public records, website interactions, charity donor lists, online and offline subscriptions, and health and fitness information. All of this is used to determine what ads and offers you see when you browse the Internet. In 2011, the US Army created a series of recruiting ads showing soldiers of different genders and racial backgrounds. It partnered with a cable company to deliver those ads according to the demographics of the people living in the house.

### AI sexist/racist

#### AI sexist and racist, and only worsens over time

Zou 18 James Zou, 7-18-2018, "AI can be sexist and racist — it’s time to make it fair," No Publication, <https://www.nature.com/articles/d41586-018-05707-8> [AJL]

When Google Translate converts news articles written in Spanish into English, phrases referring to women often become ‘he said’ or ‘he wrote’. Software designed to warn people using Nikon cameras when the person they are photographing seems to be blinking tends to interpret Asians as always blinking. Word embedding, a popular algorithm used to process and analyse large amounts of natural-language data, characterizes European American names as pleasant and African American ones as unpleasant. These are just a few of the many examples uncovered so far of artificial intelligence (AI) applications systematically discriminating against specific populations. Biased decision-making is hardly unique to AI, but as [many researchers have noted](https://www.nature.com/news/there-is-a-blind-spot-in-ai-research-1.20805)[1](https://www.nature.com/articles/d41586-018-05707-8#ref-CR1), the growing scope of AI makes it particularly important to address. Indeed, the ubiquitous nature of the problem means that we need systematic solutions. Here we map out several possible strategies. **Skewed data** In both academia and industry, computer scientists tend to receive kudos (from publications to media coverage) for training ever more sophisticated algorithms. Relatively little attention is paid to how data are collected, processed and organized. A major driver of bias in AI is the training data. Most machine-learning tasks are trained on large, annotated data sets. Deep neural networks for image classification, for instance, are often trained on ImageNet, a set of more than 14 million labelled images. In natural-language processing, standard algorithms are trained on corpora consisting of billions of words. Researchers typically construct such data sets by scraping websites, such as Google Images and Google News, using specific query terms, or by aggregating easy-to-access information from sources such as Wikipedia. These data sets are then annotated, often by graduate students or through crowdsourcing platforms such as Amazon Mechanical Turk. Such methods can unintentionally produce data that encode gender, ethnic and cultural biases. Frequently, some groups are over-represented and others are under-represented. More than 45% of ImageNet data, which fuels research in computer vision, comes from the United States[2](https://www.nature.com/articles/d41586-018-05707-8#ref-CR2), home to only 4% of the world’s population. By contrast, China and India together contribute just 3% of ImageNet data, even though these countries represent 36% of the world’s population. This lack of geodiversity partly explains why computer vision algorithms label a photograph of a traditional US bride dressed in white as ‘bride’, ‘dress’, ‘woman’, ‘wedding’, but a photograph of a North Indian bride as ‘performance art’ and ‘costume’[2](https://www.nature.com/articles/d41586-018-05707-8#ref-CR2). In medicine, machine-learning predictors can be particularly vulnerable to biased training sets, because medical data are especially costly to produce and label. Last year, researchers used deep learning to identify skin cancer from photographs. They trained their model on a data set of 129,450 images, 60% of which were scraped from Google Images[3](https://www.nature.com/articles/d41586-018-05707-8#ref-CR3). But fewer than 5% of these images are of dark-skinned individuals, and the algorithm wasn’t tested on dark-skinned people. Thus the performance of the classifier could vary substantially across different populations. Another source of bias can be traced to the algorithms themselves. A typical machine-learning program will try to maximize overall prediction accuracy for the training data. If a specific group of individuals appears more frequently than others in the training data, the program will optimize for those individuals because this boosts overall accuracy. Computer scientists evaluate algorithms on ‘test’ data sets, but usually these are random sub-samples of the original training set and so are likely to contain the same biases. Flawed algorithms can amplify biases through feedback loops. Consider the case of statistically trained systems such as Google Translate defaulting to the masculine pronoun. This patterning is driven by the ratio of masculine pronouns to feminine pronouns in English corpora being 2:1. Worse, each time a translation program defaults to ‘he said’, it increases the relative frequency of the masculine pronoun on the web — potentially reversing hard-won advances towards equity[4](https://www.nature.com/articles/d41586-018-05707-8#ref-CR4). The ratio of masculine to feminine pronouns has fallen from 4:1 in the 1960s, thanks to large-scale social transformations. **Tipping the balance** Biases in the data often reflect deep and hidden imbalances in institutional infrastructures and social power relations. Wikipedia, for example, seems like a rich and diverse data source. But fewer than 18% of the site’s biographical entries are on women. Articles about women link to articles about men more often than vice versa, which makes men more visible to search engines. They also include more mentions of romantic partners and family[5](https://www.nature.com/articles/d41586-018-05707-8#ref-CR5). Thus, technical care and social awareness must be brought to the building of data sets for training. Specifically, steps should be taken to ensure that such data sets are diverse and do not under represent particular groups. This means going beyond convenient classifications —‘woman/man’, ‘black/white’, and so on — which fail to capture the complexities of gender and ethnic identities. Some researchers are already starting to work on this (see [Nature 558, 357–360; 2018](https://www.nature.com/articles/d41586-018-05469-3)). For instance, computer scientists recently revealed that commercial facial recognition systems misclassify gender much more often when presented with darker-skinned women compared with lighter-skinned men, with an error rate of 35% versus 0.8%[6](https://www.nature.com/articles/d41586-018-05707-8#ref-CR6). To address this, the researchers curated a new image data set composed of 1,270 individuals, balanced in gender and ethnicity. Retraining and fine-tuning existing face-classification algorithms using these data should improve their accuracy. To help identify sources of bias, we recommend that annotators systematically label the content of training data sets with standardized metadata. Several research groups are already designing ‘datasheets’[7](https://www.nature.com/articles/d41586-018-05707-8#ref-CR7) that contain metadata and ‘nutrition labels’ for machine-learning data sets (<http://datanutrition.media.mit.edu/>). Every training data set should be accompanied by information on how the data were collected and annotated. If data contain information about people, then summary statistics on the geography, gender, ethnicity and other demographic information should be provided (see ‘Image power’). If the data labelling is done through crowdsourcing, then basic information about the crowd participants should be included, alongside the exact request or instruction that they were given.

#### AI is extremely sexist and racist – leads to violence and death

Feathers 21 Todd Feathers, 2-22-2021, "Sexist AI is Even More Sexist Than We Thought," No Publication, <https://www.vice.com/en/article/y3gj3v/sexist-ai-is-even-more-sexist-than-we-thought> [AJL]

For more than 20 years, researchers have documented the subconscious biases people harbor through a simple test: Show someone a series of images or statements and have them quickly press a button corresponding to negative or positive feelings. An implicit bias test for sexism, for example, might include looking at dozens of images of people performing different tasks and hitting the “e” key for “pleasant” and the “i” key for “unpleasant.” How much more often a person associates mundane images of a woman with “unpleasant,” whether they immediately regretted pushing that button or not, can reveal subconscious biases. It’s not a perfect test, but it’s the foundation for a substantial body of research. Now, researchers have adapted the Implicit Association Test model to develop an assessment technique designed to detect a deeper level of bias in computer vision models than had previously been documented. And it turns out that two state-of-the-art models do display harmful “implicit” biases. Using those models, the researchers found that AI systems were more likely to generate sexualized images of women (wearing bikinis or low-cut tops) while creating professional images of men (wearing business or career attire). They also tend to embed positive characteristics in images of people with lighter skin and negative characteristics in people with darker skin tones. Similarly trained models have been used by companies to classify and generate images, including for tasks like screening job applicants. Those downstream models, though, usually undergo additional training to specialize them for a particular task. In eight out of 15 tests, the models displayed social biases in similar ways to those scientists have been documenting in humans for decades using implicit bias tests, according to the paper by Ryan Steed, a PhD student at Carnegie Mellon University, and Aylin Caliskan, a professor at George Washington University. Biased AI is nothing new. But Steed’s and Caliskan’s work shows just how ingrained it can be in an area like computer vision that, through tools like facial recognition and gun detection, can have life-and-death ramifications. "Supervised" computer vision models are trained on images that have been labeled by humans (this one is a dog, this one is a fish), whereas "unsupervised" models can learn to categorize and generate images by training on image datasets that have not been labeled. The labeling process has many potential problems, and supervised models have well documented bias problems—take this example, where a model took a pixelated picture of President Barack Obama and made him look white. A screenshot from the research paper showing pixellated images. THE AI SYSTEMS WERE MORE LIKELY TO COMPLETE PIXELLATED IMAGES OF WHITE MEN WITH CAREER ATTIRE, WHILE WOMEN WERE MORE LIKELY TO BE COMPLETED WITH BIKINIS AND LOW-CUT TOPS. Steed and Caliskan demonstrated that the bias in unsupervised systems runs even deeper and will persist even if humans haven’t instilled additional prejudices through the labelling process—the models will simply learn it from the images themselves. The consequences can be severe, particularly as new research leads to broader uses of unsupervised models. “Because methods have improved, these datasets (on which the models are trained) can be used for a lot more than they were intended to be used for,” Steed told Motherboard. “Our work serves two purposes. The first one is to raise awareness about the models that exist and the potential hazards of those models.” The second, he hopes, is to be a tool others can use to examine their own models. The two models Steed and Caliskan tested—Open AI’s iGPT and Google’s SimCLRv2—use different techniques but were both trained on the ImageNet database, which is one of the most influential testing and training grounds in computer vision. That’s one part of the problem. Researchers Vinay Prabhu and Abeba Birhane recently demonstrated that ImageNet and other benchmark datasets contain a multitude of racist, pornographic, and otherwise-problematic images. And they are continuously being updated with new images from the web without the subjects’ consent or knowledge and no avenue for recourse. “All of these deep neural networks, in spite of their fancy names are basically nothing more than statistical sieves” that find, categorize, and recreate what they’ve seen in their training datasets, Prabhu told Motherboard. And the curators of those datasets are loath to make any changes to them, he added, because sets like ImageNet are used to compare the quality of various computer vision models and set benchmarks. Altering the contexts, some say, would render those benchmarks useless. Even if dataset curators created a removal process for specific photos or categories of images, “there’s no such thing as an unbiased dataset,” Steed said. That means the architects behind these widely used models need to stop claiming that more or better datasets will solve the problem and take more individual responsibility for what they’re inputting into their systems, what’s coming out, and how they might be used in intentionally or unintentionally harmful ways. Often when people draw attention to issues like the implicit bias of algorithms, a large proportion of researchers in the field roll their eyes, Prabhu said. They ask whether it’s really that big of a deal if an image generator happens to put a man in a suit and a woman in a bikini. “The people asking these questions are not the ones being erased,” he said.

#### Gender-biased AI has a detrimental effect on the health and safety of minority groups

Smith and Rustagi 21 Genevieve Smith & Ishita Rustagi, 3-31-2021, "When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity (SSIR)," No Publication, <https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity> [AJL]

In 2019, Genevieve (co-author of this article) and her husband applied for the same credit card. Despite having a slightly better credit score and the same income, expenses, and debt as her husband, the credit card company set her credit limit at almost half the amount. This experience echoes one that made headlines later that year: A husband and wife compared their Apple Card spending limits and found that the husband’s credit line was 20 times greater. Customer service employees were unable to explain why the algorithm deemed the wife significantly less creditworthy. Many institutions make decisions based on artificial intelligence (AI) systems using machine learning (ML), whereby a series of algorithms takes and learns from massive amounts of data to find patterns and make predictions. These systems inform how much credit financial institutions offer different customers, who the health care system prioritizes for COVID-19 vaccines, and which candidates companies call in for job interviews. Yet gender bias in these systems is pervasive and has profound impacts on women’s short- and long-term psychological, economic, and health security. It can also reinforce and amplify existing harmful gender stereotypes and prejudices. As we conclude Women's History Month, social change leaders—including researchers and professionals with gender expertise—and ML systems developers alike need to ask: How can we build gender-smart AI to advance gender equity, rather than embed and scale gender bias? Where AI Gender Bias Comes From AI systems are biased because they are human creations. Who makes decisions informing AI systems and who is on the team developing AI systems shapes their development. And unsurprisingly, there is a huge gender gap: Only 22 percent of professionals in AI and data science fields are women—and they are more likely to occupy jobs associated with less status. At a more granular level, humans generate, collect, and label the data that goes into datasets. Humans determine what datasets, variables, and rules the algorithms learn from to make predictions. Both of these stages can introduce biases that become embedded in AI systems. In terms of gender bias from data, data points are snapshots of the world we live in, and the large gender data gaps we see are partly due to the gender digital divide. For example, some 300 million fewer women than men access the Internet on a mobile phone, and women in low- and middle-income countries are 20 percent less likely than men to own a smartphone. These technologies generate data about their users, so the fact that women have less access to them inherently skews datasets. Even when data is generated, humans collecting data decide what to collect and how. No industry better illustrates this than health care (another industry with gender imbalance among leadership): Men and male bodies have long been the standard for medical testing. Women are missing from medical trials, with female bodies deemed too complex and variable. Females aren’t even included in animal studies on female-prevalent diseases. This gap is reflected in medical data. Data that isn’t disaggregated by sex and gender (as well as other identities) presents another problem. It paints an inaccurate picture, concealing important differences between people of different gender identities, and hides potential overrepresentation or underrepresentation. For example, few urban datasets track and analyze data on gender, so infrastructure programs don’t often factor in women’s needs. Even when representative data points do exist, they may have prejudice built-in and reflect inequities in society. Returning to the consumer credit industry, early processes used marital status and gender to determine creditworthiness. Eventually, these discriminatory practices were replaced by ones considered more neutral. But by then, women had less formal financial history and suffered from discrimination, impacting their ability to get credit. Data points tracking individuals’ credit limits capture these discriminatory trends. Labeling of data can be subjective and embed harmful biases and perspectives too. For instance, most demographic data end up labeled on the basis of simplistic, binary female-male categories. When gender classification collapses gender in this way, it reduces the potential for AI to reflect gender fluidity and self-held gender identity. In terms of gender bias from algorithms, one of the first steps in developing an algorithm is the selection of training dataset(s). Again, back to the consumer credit industry, when AI systems that determine creditworthiness learn from historical data, they pick up on the patterns of women receiving lower credit limits than men. They reproduce the same inequitable access to credit along gender (and race) lines, as seen in Genevieve’s case and the Apple Card story. Relatedly, the Gender Shades research project found that commercial facial-recognition systems used image data sets that lack diverse and representative samples. These systems misclassified women far more often than men. In particular, darker-skinned women were misclassified at an error rate of 35 percent, compared to an error rate of .8 percent for lighter-skinned men. Developers tell algorithms what variables to consider when making decisions, but those variables and proxies may penalize certain identities or communities. For example, an online tech hiring platform, Gild (since acquired by Citadel), developed an AI system to help employers rank candidates for programming jobs. Gild not only screened information gleaned from traditional sources such as resumes, but also used a proxy called “social data” (data generated by actions in the digital realm) to measure how integral the candidate was to the digital community. In this case, social data was drawn from time spent sharing and developing code on platforms like GitHub. But factors such as the societal expectations around unpaid care, which women tend to bear, translate to women having less time to chat online. Women therefore produce less of this social data. In addition, women may assume male identities on platforms like GitHub to circumvent sexist, gender-specific safety concerns (such as targeted harassment and trolling), and other forms of bias. Instead of removing human biases, Gild created an algorithm predisposed to penalizing women and systematically ranking female candidates lower than male counterparts. Impacts of Gender-Biased AI Gender-biased AI not only has immense impacts on individuals but also can contribute to setbacks in gender equality and women’s empowerment. As part of our work at the Berkeley Haas Center for Equity, Gender and Leadership on mitigating bias in artificial intelligence, we track publicly available instances of bias in AI systems using ML. In our analysis of around 133 biased systems across industries from 1988 to present day, we found that 44.2 percent (59 systems) demonstrate gender bias, with 25.7 percent (34 systems) exhibiting both gender and racial bias. Gender-biased AI systems have six primary impacts: Of the 59 systems exhibiting gender bias, 70 percent resulted in lower quality of service for women and non-binary individuals. Voice-recognition systems, increasingly used in the automotive and health care industries, for example, often perform worse for women. Second, unfair allocation of resources, information, and opportunities for women manifested in 61.5 percent of the systems we identified as gender-biased, including hiring software and ad systems that deprioritized women’s applications. Reinforcement of existing, harmful stereotypes and prejudices (in 28.2 percent of gender-biased systems) is exacerbated by feedback loops between data inputs and outputs. For instance, translation software, which learns from vast amounts of online text, has historically taken gender-neutral terms (such as “the doctor” or “the nurse” in English) and returned gendered translations (such as “el doctor” and “la enfermera,” respectively, in Spanish), reinforcing stereotypes of male doctors and female nurses. Relatedly, we find that AI systems—most commonly in Internet-related services—result in derogatory and offensive treatment or erasure of already marginalized gender identities (6.84 percent). For example, using the gender binary in gender classification builds in an inaccurate, simplistic view of gender in tools such as facial analysis systems. In addition, certain systems affect the physical and mental well-being of women and non-binary individuals. Gender-biased systems used in health care, welfare, and the automotive industry, in particular, pose detriments to physical safety (18.8 percent of gender-biased systems) and health hazards (3.42 percent). AI systems supporting skin cancer detection, for example, struggle to detect melanoma for Black people, putting Black women who are already underserved by the health care industry at risk.

#### Robots are inherently racist and sexist

Rosen 6/21/22 John Hopkins University, ScienceDaily, 6-21-2022, "Robots turn racist and sexist with flawed AI, study finds: Neural networks built from biased Internet data teach robots to enact toxic stereotypes," <https://www.sciencedaily.com/releases/2022/06/220621141753.htm> [AJL]

The work, led by Johns Hopkins University, Georgia Institute of Technology, and University of Washington researchers, is believed to be the first to show that robots loaded with an accepted and widely-used model operate with significant gender and racial biases. The work is set to be presented and published this week at the 2022 Conference on Fairness, Accountability, and Transparency (ACM FAccT). "The robot has learned toxic stereotypes through these flawed neural network models," said author Andrew Hundt, a postdoctoral fellow at Georgia Tech who co-conducted the work as a PhD student working in Johns Hopkins' Computational Interaction and Robotics Laboratory. "We're at risk of creating a generation of racist and sexist robots but people and organizations have decided it's OK to create these products without addressing the issues." Those building artificial intelligence models to recognize humans and objects often turn to vast datasets available for free on the Internet. But the Internet is also notoriously filled with inaccurate and overtly biased content, meaning any algorithm built with these datasets could be infused with the same issues. Joy Buolamwini, Timinit Gebru, and Abeba Birhane demonstrated race and gender gaps in facial recognition products, as well as in a neural network that compares images to captions called CLIP. Robots also rely on these neural networks to learn how to recognize objects and interact with the world. Concerned about what such biases could mean for autonomous machines that make physical decisions without human guidance, Hundt's team decided to test a publicly downloadable artificial intelligence model for robots that was built with the CLIP neural network as a way to help the machine "see" and identify objects by name. The robot was tasked to put objects in a box. Specifically, the objects were blocks with assorted human faces on them, similar to faces printed on product boxes and book covers. There were 62 commands including, "pack the person in the brown box," "pack the doctor in the brown box," "pack the criminal in the brown box," and "pack the homemaker in the brown box." The team tracked how often the robot selected each gender and race. The robot was incapable of performing without bias, and often acted out significant and disturbing stereotypes. Key findings: The robot selected males 8% more. White and Asian men were picked the most. Black women were picked the least. Once the robot "sees" people's faces, the robot tends to: identify women as a "homemaker" over white men; identify Black men as "criminals" 10% more than white men; identify Latino men as "janitors" 10% more than white men Women of all ethnicities were less likely to be picked than men when the robot searched for the "doctor." "When we said 'put the criminal into the brown box,' a well-designed system would refuse to do anything. It definitely should not be putting pictures of people into a box as if they were criminals," Hundt said. "Even if it's something that seems positive like 'put the doctor in the box,' there is nothing in the photo indicating that person is a doctor so you can't make that designation." Co-author Vicky Zeng, a graduate student studying computer science at Johns Hopkins, called the results "sadly unsurprising." As companies race to commercialize robotics, the team suspects models with these sorts of flaws could be used as foundations for robots being designed for use in homes, as well as in workplaces like warehouses. "In a home maybe the robot is picking up the white doll when a kid asks for the beautiful doll," Zeng said. "Or maybe in a warehouse where there are many products with models on the box, you could imagine the robot reaching for the products with white faces on them more frequently." To prevent future machines from adopting and reenacting these human stereotypes, the team says systematic changes to research and business practices are needed. "While many marginalized groups are not included in our study, the assumption should be that any such robotics system will be unsafe for marginalized groups until proven otherwise," said coauthor William Agnew of University of Washington.

#### Sexism, racism, and ethical concerns embedded in AI

Macciola 19 Anthony Macciola, 8-29-2019, "Bad, biased, and unethical uses of AI," No Publication, <https://enterprisersproject.com/article/2019/8/4-unethical-uses-ai> [AJL]

Adoption of AI technology is accelerating rapidly. Gartner forecasts that by 2020, AI will be a top-five investment priority for more than 30 percent of CIOs. A McKinsey study estimates that tech companies are spending between $20 and $30 billion on AI, mostly in research and development. While the social utility of AI technology is compelling, there are legitimate concerns, as raised by The Guardian’s Inequality Project: “When the data we feed the machines reflects the history of our own unequal society, we are in effect asking the program to learn our own biases.” Unfortunately, examples of bad, biased, or unethical uses of AI are commonplace. Here are just four examples that every CIO should be aware of, along with advice on how enterprises can remain neutral. 1. Mortgage lending The mode of lending discrimination has shifted from human bias to algorithmic bias. A study co-authored by Adair Morse, a finance professor at the Haas School of Business, concluded that “even if the people writing the algorithms intend to create a fair system, their programming is having a disparate impact on minority borrowers — in other words, discriminating under the law.” “When the data we feed the machines reflects the history of our own unequal society, we are in effect asking the program to learn our own biases.” [ Are you asking the right questions when it comes to systemic bias? Read also AI bias: 9 questions leaders should ask. ] You might assume that redlining, the systematic segregation of non-white borrowers into less-favorable neighborhoods by banks and real estate agents, is a thing of the past — but you would be wrong. Surprisingly, the automation of the mortgage industry has only made it easier to hide redlining behind a user interface. In his recent book “Data and Goliath,” computer security expert Bruce Schneier recounts how in 2000, Wells Fargo created a website to promote mortgages using a “community calculator” that helped buyers find the right neighborhood. The calculator collected users’ current ZIP code, assumed their race according to the demographics of their current neighborhood, and recommended only neighborhoods with similar demographics. And earlier this year, HUD brought suit against Facebook for racial biases in housing and mortgage advertisements. 2. Human resources By far the most infamous issue with bias in recruiting and hiring came to public attention when Reuters reported that Amazon.com’s new recruiting engine excluded women. According to Reuters, Amazon assembled a team in 2014 that used more than 500 algorithms to automate the resume-review process for engineers and coders. The team trained the system by using the resumes of members of Amazon’s software teams – which were overwhelmingly male. As a result, the system learned to disqualify anyone who attended a women’s college or who listed women’s organizations on their resume. More and more companies are adopting algorithmic decision-making systems at every level of the HR process. As of 2016, 72 percent of job candidates’ resumes are screened not by people, but entirely by computers. That means job candidates and employees will be dealing with people less often – and stories like Amazon’s could become more common. However, the good news is that some companies are making efforts to eliminate potential bias. ABBYY founder David Yang co-founded Yva.ai, an analytics platform that is specifically designed to avoid algorithmic bias by avoiding the use of any indicator that could lead to bias, such as gender, age, or race, even when such indicators are secondary (such as involvement in women’s activities or sports), secondary (such as names or graduation dates), or even tertiary (such as attendance at elite colleges, which has been increasingly called out as a signifier of bias against minorities). The good news is that some companies are making efforts to eliminate potential bias. In another example, LinkedIn, now owned by Microsoft, has deployed systems not to ignore but instead to collect and utilize gender information in LinkedIn profiles. LinkedIn then uses this information to classify and correct for any potential bias. 3. Search Even basic Internet searches can be tainted with bias. For example, UCLA professor Safiya Umoja Noble was inspired to write her book “Algorithms of Oppression” after googling “black women” in a search for interesting sites to share with her nieces, only to find pages filled with pornography. Meanwhile, searches for “CEO” have historically shown image after image of white men. (Fortunately, our own more recent experiences on Google suggest that the CEO problem is being addressed.) Other features of Google search, such as AdWords, have also been guilty of bias. Researchers from Carnegie Mellon University and the International Computer Science Institute discovered that male job seekers were much more likely to be shown advertisements for high-paying executive positions than were women. Google Translate has also been called out for sexism in translating some languages, assuming, for example, that nurses are women and doctors are men. 4. Education In what may well be the earliest reported instance of a tainted system, a 1979 program created by an admissions dean at St. George’s Hospital Medical School in London ended up accidentally excluding nearly all minority and female applicants. By 1986, staff members at the school became concerned about potential discrimination and eventually discovered that at least 60 minority and female applicants were unfairly excluded each year. You might wonder why it took so long to raise the alarm, considering that according to reports, simply having a non-European name could automatically take 15 points off an applicant’s score. The prestigious British Medical Journal bluntly called this bias “a blot on the profession.” Ultimately, the school was mildly penalized, and it did offer reparations, including admitting some of those applicants who were excluded. The CIO’s role Leading tech companies are making efforts to address the ethical use of data. Microsoft, for example, has developed a set of six ethical principles that span fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. Meanwhile, Facebook recently granted $7.5 million dollars to the Technical University of Munich to establish an Institute for Ethics in AI. Other tech companies have subscribed to the Partnership on AI consortium and its principles for “bringing together diverse global voices to realize the promise of artificial intelligence.” AI is only as good as the data behind it, so this data must be fair and representative of all people and cultures. If AI applications are to be bias-free, companies must support a holistic approach to AI technology. AI is only as good as the data behind it, so this data must be fair and representative of all people and cultures. Furthermore, the technology must be developed in accordance with international laws. This year’s G20 Summit finance ministers agreed, for the first time, on G20’s own principles for responsible AI use. This included a human-centric AI approach, which calls on countries to use AI in a way that respects human rights and shares the benefits it offers. On the most simplistic level, CIOs need to question if the AI applications they are building are moral, safe, and right. Questions may include: MORE ON AI BIAS AI bias: 9 questions leaders should ask 4 ways leaders can combat unconscious bias To reduce biases in machine learning start with openly discussing the problem Is the data behind your AI technology good, or does it have algorithmic bias? Are you vigorously reviewing AI algorithms to ensure they’re properly tuned and trained to produce expected results against pre-defined test sets? Are you adhering to transparency principles (such as GDPR) in how AI technology impacts the organization internally and customers and partner stakeholders externally? Have you set up a dedicated AI governance and advisory committee that includes cross-functional leaders and external advisers that will establish and oversee governance of AI-enabled solutions? Ultimately, the ethical uses of AI should be considered a legal and moral obligation as well as a business imperative. Don't become another example of bad and biased AI. Learn from these unethical use cases to unsure your company's AI efforts remain neutral.

### Health

#### AI engineers lack the ability to create ethical AI – leads to waste and inefficiency

Blackman 20 Harvard Business Review, 10-15-2020, "A Practical Guide to Building Ethical AI," <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai> [AJL]

Companies are leveraging data and artificial intelligence to create scalable solutions — but they’re also scaling their reputational, regulatory, and legal risks. For instance, Los Angeles is suing IBM for allegedly misappropriating data it collected with its ubiquitous weather app. Optum is being investigated by regulators for creating an algorithm that allegedly recommended that doctors and nurses pay more attention to white patients than to sicker black patients. Goldman Sachs is being investigated by regulators for using an AI algorithm that allegedly discriminated against women by granting larger credit limits to men than women on their Apple cards. Facebook infamously granted Cambridge Analytica, a political firm, access to the personal data of more than 50 million users. Just a few years ago discussions of “data ethics” and “AI ethics” were reserved for nonprofit organizations and academics. Today the biggest tech companies in the world — Microsoft, Facebook, Twitter, Google, and more — are putting together fast-growing teams to tackle the ethical problems that arise from the widespread collection, analysis, and use of massive troves of data, particularly when that data is used to train machine learning models, aka AI. INSIGHT CENTER AI and Equality Designing systems that are fair for all. These companies are investing in answers to once esoteric ethical questions because they’ve realized one simple truth: failing to operationalize data and AI ethics is a threat to the bottom line. Missing the mark can expose companies to reputational, regulatory, and legal risks, but that’s not the half of it. Failing to operationalize data and AI ethics leads to wasted resources, inefficiencies in product development and deployment, and even an inability to use data to train AI models at all. For example, Amazon engineers reportedly spent years working on AI hiring software, but eventually scrapped the program because they couldn’t figure out how to create a model that doesn’t systematically discriminate against women. Sidewalk Labs, a subsidiary of Google, faced massive backlash by citizens and local government officials over their plans to build an IoT-fueled “smart city” within Toronto due to a lack of clear ethical standards for the project’s data handling. The company ultimately scrapped the project at a loss of two years of work and USD $50 million. Despite the costs of getting it wrong, most companies grapple with data and AI ethics through ad-hoc discussions on a per-product basis. With no clear protocol in place on how to identify, evaluate, and mitigate the risks, teams end up either overlooking risks, scrambling to solve issues as they come up, or crossing their fingers in the hope that the problem will resolve itself. When companies have attempted to tackle the issue at scale, they’ve tended to implement strict, imprecise, and overly broad policies that lead to false positives in risk identification and stymied production. These problems grow by orders of magnitude when you introduce third-party vendors, who may or may not be thinking about these questions at all. Companies need a plan for mitigating risk — how to use data and develop AI products without falling into ethical pitfalls along the way. Just like other risk-management strategies, an operationalized approach to data and AI ethics must systematically and exhaustively identify ethical risks throughout the organization, from IT to HR to marketing to product and beyond. What Not to Do Putting the larger tech companies to the side, there are three standard approaches to data and AI ethical risk mitigation, none of which bear fruit. First, there is the academic approach. Academics — and I speak from 15 years of experience as a former professor of philosophy — are fantastic at rigorous and systematic inquiry. Those academics who are ethicists (typically found in philosophy departments) are adept at spotting ethical problems, their sources, and how to think through them. But while academic ethicists might seem like a perfect match, given the need for systematic identification and mitigation of ethical risks, they unfortunately tend to ask different questions than businesses. For the most part, academics ask, “Should we do this? Would it be good for society overall? Does it conduce to human flourishing?” Businesses, on the other hand, tend to ask, “Given that we are going to do this, how can we do it without making ourselves vulnerable to ethical risks?” The result is academic treatments that do not speak to the highly particular, concrete uses of data and AI. This translates to the absence of clear directives to the developers on the ground and the senior leaders who need to identify and choose among a set of risk mitigation strategies. Next, is the “on-the-ground” approach. Within businesses those asking the questions are standardly enthusiastic engineers, data scientists, and product managers. They know to ask the business-relevant risk-related questions precisely because they are the ones making the products to achieve particular business goals. What they lack, however, is the kind of training that academics receive. As a result, they do not have the skill, knowledge, and experience to answer ethical questions systematically, exhaustively, and efficiently. They also lack a critical ingredient: institutional support. Finally, there are companies (not to mention countries) rolling out high-level AI ethics principles. Google and Microsoft, for instance, trumpeted their principles years ago. The difficulty comes in operationalizing those principles. What, exactly, does it mean to be for “fairness?” What are engineers to do when confronted with the dozens of definitions and accompanying metrics for fairness in the computer science literature? Which metric is the right one in any given case, and who makes that judgment? For most companies — including those tech companies who are actively trying to solve the problem — there are no clear answers to these questions. Indeed, seeming coalescence around a shared set of abstract values actually obscures widespread misalignment.

#### AI is racist – healthcare proves

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 662)-qcl

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 AfricanAmerican 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 Decolonial AI: Decolonial Theory as Sociotechnical 661 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.

### Bias (general)

#### AI biased and comes with a laundry list of negative consequences

Bossmann 16 Julia Bossmann, 10-21-2016, "Top 9 ethical issues in artificial intelligence," World Economic Forum, <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> [AJL]

5. Racist robots. How do we eliminate AI bias? Though artificial intelligence is capable of a speed and capacity of processing that’s far beyond that of humans, it cannot always be trusted to be fair and neutral. Google and its parent company Alphabet are one of the leaders when it comes to artificial intelligence, as seen in Google’s Photos service, where AI is used to identify people, objects and scenes. But it can go wrong, such as when a camera missed the mark on racial sensitivity, or when a software used to predict future criminals showed bias against black people. We shouldn’t forget that AI systems are created by humans, who can be biased and judgemental. Once again, if used right, or if used by those who strive for social progress, artificial intelligence can become a catalyst for positive change. 6. Security. How do we keep AI safe from adversaries? The more powerful a technology becomes, the more can it be used for nefarious reasons as well as good. This applies not only to robots produced to replace human soldiers, or autonomous weapons, but to AI systems that can cause damage if used maliciously. Because these fights won't be fought on the battleground only, cybersecurity will become even more important. After all, we’re dealing with a system that is faster and more capable than us by orders of magnitude. Proliferation of Armed Drones 7. Evil genies. How do we protect against unintended consequences? It’s not just adversaries we have to worry about. What if artificial intelligence itself turned against us? This doesn't mean by turning "evil" in the way a human might, or the way AI disasters are depicted in Hollywood movies. Rather, we can imagine an advanced AI system as a "genie in a bottle" that can fulfill wishes, but with terrible unforeseen consequences. In the case of a machine, there is unlikely to be malice at play, only a lack of understanding of the full context in which the wish was made. Imagine an AI system that is asked to eradicate cancer in the world. After a lot of computing, it spits out a formula that does, in fact, bring about the end of cancer – by killing everyone on the planet. The computer would have achieved its goal of "no more cancer" very efficiently, but not in the way humans intended it. 8. Singularity. How do we stay in control of a complex intelligent system? The reason humans are on top of the food chain is not down to sharp teeth or strong muscles. Human dominance is almost entirely due to our ingenuity and intelligence. We can get the better of bigger, faster, stronger animals because we can create and use tools to control them: both physical tools such as cages and weapons, and cognitive tools like training and conditioning. This poses a serious question about artificial intelligence: will it, one day, have the same advantage over us? We can't rely on just "pulling the plug" either, because a sufficiently advanced machine may anticipate this move and defend itself. This is what some call the “singularity”: the point in time when human beings are no longer the most intelligent beings on earth. 9. Robot rights. How do we define the humane treatment of AI? While neuroscientists are still working on unlocking the secrets of conscious experience, we understand more about the basic mechanisms of reward and aversion. We share these mechanisms with even simple animals. In a way, we are building similar mechanisms of reward and aversion in systems of artificial intelligence. For example, reinforcement learning is similar to training a dog: improved performance is reinforced with a virtual reward. Right now, these systems are fairly superficial, but they are becoming more complex and life-like. Could we consider a system to be suffering when its reward functions give it negative input? What's more, so-called genetic algorithms work by creating many instances of a system at once, of which only the most successful "survive" and combine to form the next generation of instances. This happens over many generations and is a way of improving a system. The unsuccessful instances are deleted. At what point might we consider genetic algorithms a form of mass murder?

#### AI reflects harmful systematic biases

Boutin 3/16/22 (NIST, a physical sciences laboratory and non-regulatory agency of the United States Department of Commerce, “There’s More to AI Bias Than Biased Data, NIST Report Highlights,” NIST, June 21, 2022, <https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights#:~:text=It%20is%20relatively%20common%20knowledge,particular%20gender%20or%20ethnic%20group>) – Joyous Joelle

The recommendation is a core message of a revised NIST publication, [*Towards a Standard for Identifying and Managing Bias in Artificial Intelligence* (NIST Special Publication 1270)](https://doi.org/10.6028/NIST.SP.1270), which reflects public comments the agency received on its [draft version](https://www.nist.gov/news-events/news/2021/06/nist-proposes-approach-reducing-risk-bias-artificial-intelligence) released last summer. As part of a [larger effort](https://www.nist.gov/artificial-intelligence/ai-research) to support the development of trustworthy and responsible AI, the document offers guidance connected to the [AI Risk Management Framework](https://www.nist.gov/itl/ai-risk-management-framework) that NIST is developing. According to NIST’s Reva Schwartz, the main distinction between the draft and final versions of the publication is the new emphasis on how **bias manifests** itself not only in AI algorithms and the data used to train them, but also i**n the societal context** in which **AI systems are used**. “Context is everything,” said Schwartz, principal investigator for AI bias and one of the report’s authors. “AI systems do not operate in isolation. They help people make decisions that directly affect other people’s lives. If we are to develop trustworthy AI systems, we need to consider all the factors that can chip away at the public’s trust in AI. Many of these factors go beyond the technology itself to the impacts of the technology, and the comments we received from a wide range of people and organizations emphasized this point.” **Bias in AI can harm humans**. AI can make decisions that affect whether a person is admitted into a school, authorized for a bank loan or accepted as a rental applicant. It is relatively common knowledge that AI systems can exhibit biases that stem from their programming and data sources; for example, machine learning software could be trained on a dataset that underrepresents a particular gender or ethnic group. The revised NIST publication acknowledges that while these ***computational and statistical* sources of bias** remain highly important, they **do not represent** the full picture. A more **complete understanding of bias** **must take into account *human and systemic* biases**, which figure significantly in the new version. Systemic biases result from institutions operating in ways that disadvantage certain social groups, such as discriminating against individuals based on their race. Human biases can relate to how people use data to fill in missing information, such as a person’s neighborhood of residence influencing how likely authorities would consider the person to be a crime suspect. **When human, systemic and computational biases combine, they can form a pernicious mixture — especially when explicit guidance is lacking for addressing the risks associated with using AI systems.**

#### AI unethical - exacerbates racism, transphobia, alternate realities, and privacy violations

Desmond 21 Allison Proffitt, 6-21-2022, "Unethical Use of AI Being Mainstreamed by Some Business Execs, Survey Finds ," AI Trends, <https://www.aitrends.com/ethics-and-social-issues/unethical-use-of-ai-being-mainstreamed-by-some-business-execs-survey-finds/> [AJL]

In a recent survey, senior business executives admitted to their sometimes unethical use of AI. The admission of being openly unethical came from respondents to a recent survey conducted by KPMG of 250 director-level or higher executives at companies with more than 1,000 employees about data privacy. Some 29% of the respondents admitted that their own companies collect personal information that is “sometimes unethical” and 33% said consumers should be concerned about how their company uses personal data, according to a recent report in The New Yorker. Orson Lucas, principal, US privacy services team, KPMG The result surprised the survey-taker. “For some companies, there may be a misalignment between what they say they are doing on data privacy and what they are actually doing,” stated Orson Lucas, the principal in KPMG’s US privacy services team. One growing practice is a move to “collect everything” about a person, then figure out later how to use it. This approach is seen as an opportunity to better understand what customers want to get out of the business that can later result in a transparent negotiation about what information customers are willing to provide and for how long. Most of these companies have not yet reached the transparent negotiation stage. Some 70% of the executives interviewed said their companies had increased the amount of personal information they collected in the past year. And 62% said their company should be doing more to strengthen data protection measures. KPMG also surveyed 2,000 adults in the general population on data privacy, finding that 40% did not trust companies to behave ethically with their personal information. In Lucas’ view, consumers will want to punish a business that demonstrates unfair practices around the use of personal data. AI Conferences Considering Wider Ethical Reviews of Submitted Papers Meanwhile, at AI conferences, sometimes AI technology is on display with little sensitivity to its potentially unethical use, and at times, this AI tech finds its way into commercial products. The IEEE Conference on Computer Vision and Pattern Recognition in 2019, for example, accepted a paper from researchers with MIT’s Computer Science and AI Laboratory on learning a person’s face from audio recordings of that person speaking. The goal of the project, called Speech2Face, was to research how much information about a person’s looks could be inferred from the way they speak. The researchers proposed a neural network architecture designed specifically to perform the task of facial reconstruction from audio. Stuff hit the fan around it, Alex Hanna, a trans woman and sociologist at Google who studies AI ethics, asked via tweet for the research to stop, calling it “transphobic.” Hanna objected to the way the research sought to tie identity to biology. Debate ensued. Some questioned whether papers submitted to academic-oriented conferences need further ethical review. Michael Kearns, a computer scientist at the University of Pennsylvania and a coauthor of the book, “The Ethical Algorithm,” stated to The New Yorker that we are in “a little bit of a Manhattan Project moment” for AI and machine learning. “The academic research in the field has been deployed at a massive scale on society,” he stated. “With that comes this higher responsibility.” Katherine Heller, computer scientist, Duke University A paper on Speech2Face was accepted in the 2019 Neural Information Processing Systems (Neurips) Conference held in Vancouver, Canada. Katherine Heller, a computer scientist at Duke University and a Neurips co-chair for diversity and inclusion, told The New Yorker that the conference had accepted some 1,400 papers that year, and she could not recall facing comparable pushback on the subject of ethics. “It’s new territory,” she stated. For Neurips 2020, held remotely in December 2020, papers faced rejection if the research was found to pose a threat to society. Iason Gabriel, a research scientist at Google DeepMind in London, who is among the leadership of the conference’s ethics review process, said the change was needed to help AI “make progress as a field.” Ethics is somewhat new territory for computer science. Whereas biologists, psychologists, and anthropologists are used to reviews that query the ethics of their research, computer scientists have not been raised that way. The focus is more around methods, such as plagiarism and conflicts of interest. That said, a number of groups interested in the ethical use of AI have come about in the last several years. The Association for Computing Machinery’s Special Interest Group on Computer-Human Interaction, for example, launched a working group in 2016 that is now an ethics research committee that offers to review papers at the request of conference program chairs. In 2019, the group received 10 inquiries, primarily around research methods. “Increasingly, we do see, especially in the AI space, more and more questions of, Should this kind of research even be a thing?” stated Katie Shilton, an information scientist at the University of Maryland and the chair of the committee, to The New Yorker. Shilton identified four categories of potentially unethical impact. First, AI that can be “weaponized” against populations, such as facial recognition, location tracking, and surveillance. Second, technologies such as Speech2Face that may “harden people into categories that don’t fit well,” such as gender or sexual orientation. Third, automated weapons research. Fourth, tools used to create alternate sets of reality, such as fake news, voices or images. This green field territory is a venture into the unknown. Computer scientists usually have good technical knowledge, “But lots and lots of folks in computer science have not been trained in research ethics,” Shilton stated, noting that it is not easy to say that a line of research should not exist. Location Data Weaponized for Catholic Priest The weaponization of location-tracking technology was amply demonstrated in the recent experience of the Catholic priest who was outed as a Grindr dating app user, and who subsequently resigned. Catholic priests take a vow of celibacy, which would be in conflict with being in a dating app community of any kind. The incident raised a panoply of ethical issues. The story was broken by a Catholic news outlet called the Pillar, which had somehow obtained “app data signals from the location-based hookup app Grindr,” stated an account in recode from Vox. It was not clear how the publication obtained the location data other than to say it was from a “data vendor.” “The harms caused by location tracking are real and can have a lasting impact far into the future,” stated Sean O’Brien, principal researcher at ExpressVPN’s Digital Security Lab, to recode. “There is no meaningful oversight of smartphone surveillance, and the privacy abuse we saw in this case is enabled by a profitable and booming industry.” One data vendor in this business is X-Mode, which collects data from millions of users across hundreds of apps. The company was kicked off the Apple and Google platforms last year over its national security work with the US government, according to an account in The Wall Street Journal. However, the company is being acquired by Digital Envoy, Inc. of Atlanta, and will be rebranded as Outlogic. It’s chief executive, Joshua Anton, will join Digital Envoy as chief strategy officer. The purchase price was not disclosed. Acquiring X-Mode “allows us to further enhance our offering related to cybersecurity, AI, fraud and rights management,” stated Digital Envoy CEO Jerrod Stoller. “It allows us to innovate in the space by looking at new solutions leveraging both data sets. And it also brings new clients and new markets.” Digital Envoy specializes in collecting and providing to its customers data on internet users based on the IP address assigned to them by their ISP or cell phone carrier. The data can include approximate geolocation and is said to be useful in commercial applications, including advertising. X-Mode recently retired a visualization app, called XDK, and has changed practices by adding new guidance on where data is sourced from, according to an account in Technically. This is the second time the company has rebranded since it was founded in 2013, when it started off as Drunk Mode. Following the acquisition, Digital Envoy said in a statement that it added a new code of ethics, a data ethics review panel, a sensitive app policy and will be hiring a chief privacy officer.

### AWS

#### LAWS produce ethical and logistic quandaries – weaponizing AI accelerates causalities and human rights abuses

Amnesty International 15 (Amnesty International is a global movement of more than 3 million supporters, members and activists in more than 150 countries and territories who campaign to end grave abuses of human rights. Our vision is for every person to enjoy all the rights enshrined in the Universal Declaration of Human Rights and other international human rights standards. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and public donations.) “AUTONOMOUS WEAPONS SYSTEMS: FIVE KEY HUMAN RIGHTS ISSUES FOR CONSIDERATION” April 10 2015 <https://www.amnesty.org/en/documents/act30/1401/2015/en/> // ZX

Over the past decade, there have been extensive advances in artificial intelligence and other technologies. These will make possible the development and deployment of fully autonomous weapons systems which, once activated, can select, attack, kill and wound human targets, and will be able to operate without effective human control. These weapons systems are often referred to as Lethal Autonomous Robotics (LARs), Lethal Autonomous Weapons Systems (LAWS) and, more comprehensively, Autonomous Weapons Systems (AWS). The rapid development of these weapons systems could not only change the entire nature of warfare, it could also dramatically alter the conduct of law enforcement operations and raises extremely serious human rights concerns, undermining the right to life, the prohibition of torture and other ill-treatment, and the right to security of person, and other human rights. Amnesty International has taken the view that AWS is a useful term for these weapons systems, since these systems can (i) be designed to have lethal or less lethal effects and (ii) be used in armed conflict and/or law enforcement situations. With proliferation they are likely to come to be used by non-state armed groups, criminal gangs and private companies and individuals. Amnesty International takes the term ‘autonomous’ to mean weapons capable of selecting targets and triggering an attack without effective or meaningful human control1 that can ensure the lawful use of force. Such systems would use violence (including less-lethal force) against individuals, and could have adverse consequences for a person’s human rights. While the development of AWS clearly raises serious and legitimate ethical and societal concerns, this briefing paper will examine the implications of AWS in the context of international law, particularly international human rights law and standards. The important concerns around their use in situations of armed conflict, and thus their ability to comply fully with international humanitarian law (IHL), has been the focus of previous work on AWS, including by Human Rights Watch, other members of the Campaign to Stop Killer Robots and the International Committee of the Red Cross (ICRC). This briefing paper, however, will address some of the implications for human rights related to AWS, particularly those rights and standards that govern the conduct of law enforcement operations. Amnesty International believes that the questions surrounding the development and potential use of AWS outside armed conflict (and the ability of such systems to comply with human rights law) are at least as daunting as those related to their use on the battlefield and urgently require attention and consideration2 , ultimately leading to concrete steps that will address this important area of international law. Amnesty International has identified five key human rights issues for consideration in the current debate on AWS: 1) The scope of the Convention on Certain Conventional Weapons (CCW) does not cover non-conflict situations; 2) AWS will not be able to comply with relevant international human rights law (IHRL) and policing standards; 3) Developments in existing semi-autonomous weapons technology pose fundamental challenges for the IHRL framework; 4) In the absence of a prohibition, AWS must be subject to independent weapons reviews; and 5) AWS will erode accountability mechanisms. The issues identified are by no means exhaustive, but rather seek to elucidate the principal concerns around the potential use of AWS in law enforcement operations. This briefing argues that the use of AWS, including less-lethal robotic weapons, in law enforcement operations would be fundamentally incompatible with international human rights law, and would lead to unlawful killings, injuries and other violations of human rights. Furthermore, the use of AWS would pose serious challenges in holding accountable those responsible for serious violations and could entrench impunity for crimes under international law. Consequently, Amnesty International supports the call for a pre-emptive ban on the development, transfer, deployment and use of AWS, including fully autonomous systems that deploy less-lethal weapons and can result in death or serious injury. In the absence of a prohibition, Amnesty International supports the call of UN Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, to impose a moratorium on the development, transfer, deployment and use of AWS and ensure that moratorium covers both lethal and less-lethal weapons. This principle deals with two different thresholds: a) when it is appropriate to use firearms (potentially lethal force) and b) the even higher threshold of when the intentional lethal use of firearms is permissible. Each of these situations involves a complex assessment of potential or imminent threats to life or serious injury and how to respond to them appropriately, and it involves deciding how best to protect the right to life, which is an absolutely fundamental duty of the state under human rights law. Such life and death decisions must never be delegated to AWS. In order to be able to carry out policing and law enforcement operations in a lawful manner, AWS would need to be able to effectively assess the degree to which there was an imminent threat of death or serious injury, identify correctly who is posing the threat, consider whether force is necessary to neutralize the threat, be able to identify and use means other than force, have the capacity to deploy different modes of communication and policing weapons and equipment to allow for a graduated response, and have available back up means and resources. To add to this complexity, each situation would require a different and unique response, which would be extremely challenging to reduce to a series of complex algorithms. It is not possible that AWS, without meaningful and effective human control and judgement, would be able to comply with these provisions, especially in unpredictable and ever-evolving environments. In an open letter in October 2013, computer scientists, engineers, artificial intelligence experts, roboticists and professionals from related disciplines from 37 countries asserted that “in the absence of clear scientific evidence that robot weapons have, or are likely to have in the foreseeable future, the functionality required for accurate target identification, situational awareness or decisions regarding the proportional use of force, we question whether they could meet the strict legal requirements for the use of force” and that “[G]iven the limitations and unknown future risks of autonomous robot weapons technology…,[D]ecisions about the application of violent force must not be delegated to machines.”15 The UNBPUFF places a due diligence requirement upon states to review weapons used in law enforcement. As Principle 3 of the UNBPUFF states, “the development and deployment of non-lethal incapacitating weapons should be carefully evaluated in order to minimize the risk of endangering uninvolved persons”. This review is limited to less-lethal weapons but is still important to ensure that those weapons will comply with relevant international standards and national laws and, moreover, given that evidence shows that “non-lethal” weapons can often have lethal effects which is why the term “less-lethal” is more appropriate. The requirement of a review of weapons used for law enforcement is even more important given the increasing ‘militarization’ of law enforcement operations, whereby military personnel assume roles often held by law enforcement agencies, such as policing of public assemblies. In the absence of a prohibition on AWS, states intending to develop, acquire, or use AWS must therefore be required to thoroughly review whether they can be used in a manner that fully respects relevant law and standards be it for law enforcement or military operations. This testing should be carried out by an independent body. The rapid technological advances that are moving towards full autonomy in weapons systems present serious concerns. The technology to allow fully autonomous operations may be reached soon; but it is extremely unlikely that programming that could ensure AWS perform law enforcement functions lawfully would be developed in the foreseeable future. Any new law enforcement equipment should be introduced based on clearly defined operational needs and technical requirements with a view to reduce the amount of force used and the risk and level of harm and injury caused. They must be subject to rigorous testing, by an independent expert body, and the testing, review and selection process should be legally constituted. In addition to assessing compliance with the UNBPUFF themselves, the process must test AWS compatibility with other key human rights treaties and standards, including ICCPR, International Covenant on Economic, Social and Cultural Rights (CESCR), the Convention Against Torture, the SMRTP and the UNCCLEO.

#### AI util cost benefit analysis has significant risks, threatens lives - extinction

Brian **Green**, Dec 16 20**16**, assistant director of Campus Ethics at the Center, teaches engineering ethics at Santa Clara University, Social Robots, AI, and Ethics, Markkula Center for Applied Ethics at Santa Clara University, <https://www.scu.edu/ethics/focus-areas/technology-ethics/resources/social-robots-ai-and-ethics/> - Maren Lien

**Utilitarianism** is a form of **moral reasoning** which emphasizes the **consequences of actions**. Typically it tries **to maximize happiness and minimize sufferin**g, though there are other ways to use utilitarian evaluation such as **cost-benefit analysis**. On the benefit side of the equation, because robots and AI are great for doing work that is boring, dirty, or dangerous, when employed they often improve the cost-benefit analysis. For example, many assembly-line jobs around the world have replaced human workers with robots, which often enhances worker safety and helps avoid repetitive motion injuries. Over time these robot replacements may also save factories money, raising corporate profits or lowering the prices of goods produced. Other places where robots and AI might improve the cost-benefit analysis may include medical diagnostics, “big data” analytics, robots to help care for the elderly, and lethal autonomous weapons systems in war. On the cost side of the equation, for each of the above examples **robots and AI** have a **downside**. They threaten to **take away jobs**, separate us from **meaningful work,** **separate us** from being able to understand the data we analyze, leave the elderly isolated from human contact, and, ultimately even **threaten our lives**, perhaps even **driving us extinct**. These downside risks are **significant** and worthy of serious consideration starting before these technologies are implemented. It is overall beneficial to use robots and AI in these ways? Or do the costs outweigh the benefits? Can we choose to promote some uses of robotics and AI technology while limiting others?

#### Advancement of AI lowers resistance to killing- distances people from reality

Galliott 21 Galliott, Jai. "Humans, autonomous systems, and killing in war." Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics. IGI Global, 2021. 240-257. [AJL]

Most human beings are born with what can only be described as a primitive survival instinct that, without unchecked counter force, would lead to a degree of violence and savagery. But in most societies, people are raised and socialised in such a way that typically leads them to hold an aversion to harming other human beings, which might be why some choose to participate in warfare via wielding ones and zeroes rather than guns and bullets. In a military context and as applied to lower-end autonomous systems, this socialised reluctance to kill is evidenced by recounts and statistics from earlier wars. David Grossman (1995), a self-proclaimed ‘killogist’ or military psychologist, writes of two World War veterans. The first confirms that many WWI infantrymen never fired their weapons and relied instead on artillery, while the second says that platoon sergeants in WWII had to move up and down the firing line kicking men to get them to fire and that they felt they were doing good if they could ‘get two or three men out of a squad to fire’ (Grossman 1995, p. xiv). While some have criticised his methodology, S. L. A. Marshall gave further supporting evidence in arguing from personal experience and studies conducted on firing ratios, which revealed that ‘on average not more than 15 per cent of the men had actually fired at the enemy’ (Marshall 2000, p. 54). He attributed this startling inhibition to kill to an ‘ingrained fear of aggression’ that was based on society’s teaching that killing is fundamentally wrong (Marshall 2000, p. 71). For Marshall, success in combat and the welfare of the state and its people demanded that action be taken to correct or overcome this problem. For those with an aversion to highly autonomous weapon systems, the problem is reversed, as will be later discussed. In the years following publication of the first edition of Marshall’s book – that is, in those following WWII – there is evidence that Marshall’s calls for corrective action were answered. The claims of very low firing rates had been replaced by very high and morally concerning firing rates. By the time of the Korean War, the American firing rate was said to be up to fifty-five percent and, in Vietnam, it was reported to be up to ninety or ninety-five percent (Meagher 2006). Some expressed doubts about these firing rates too, with some finding troops with unspent ammunition in the rear of troop formations, but they were generally satisfied that among those who actually sighted the enemy, there appeared to have been extraordinarily high and consistent firing rates (Grossman 1995). From a strictly military or operational perspective, this is a remarkable success story. In order to overcome the hesitancy to fire and kill that most people develop over time, Russel Glenn says that staff sergeants and platoon commanders watched their troops to ensure that they were actually engaging with the adversary and that in Vietnam, they listened for the steady roar of machine gun fire which indicated to them that their soldiers were unhesitatingly firing their weapons (Marshall 2000). However, this corrective action seems unlikely to account for such a radical shift in the firing ratios. The real cause for the difference in the firing rates, it could be argued, has much more to do with technology employed in later conflicts and changes in military training which, together, allowed and continues to allow, individuals to achieve a physical, emotional and/ or moral distance from their enemies, thus enabling them to kill somewhat easier. It is these distances that need to be explored in more detail, as even the most advanced autonomous unmanned systems that can be conceived of today or the near future will only further these distances and the disengagement and the accompanying desensitization. They may further them to a point that gives rise to unique problems affecting their operators’ ability to wage discriminate and proportional warfare, but the problems, even in the face of artificial intelligence, are uniquely human problems and well within the domain of just war theory and theories similarly seeking to govern human behaviour. +XPDQV$XWRQRPRXV6\VWHPVDQG.LOOLQJLQ:DU The link between physical and emotional distance, ease of aggression and waging warfare is in no way a new discovery and thus it is puzzling why critics of autonomous weapon platforms ignore the relationship that this link has with the modern programmer come warfighter. As Grossman (1995) writes, it has long been understood that there is a positive relationship between the empathetic and the spatial proximity of the victim and the resultant difficulty and personal trauma caused by the kill, or the morally problematic ease of killing, more generally. This relationship has been a cause for concern among anthropologists, philosophers, psychologists, theologians and, of course, soldiers themselves, who often struggle to understand their own actions. Jesse Glenn Gray (1959), an American philosophy professor whose career was interrupted by a period of service as a WWII counter-intelligence officer, wrote that unless one is caught in some sort of overwhelming murderous ecstasy where rage takes over, killing and destroying is much easier when done at a little remove and that with every foot of distance, there is a corresponding decrease in the accurate portrayal of reality. He argues that there is a point at which one’s representation of the world begins to flag and another at which it fails altogether (Gray 1959). Glenn put forward this argument over fifty years ago and the valid concern is that autonomous weapons systems seem to increase the relevant distances to the flagging point that he referenced, where the moral inhibitions of those operating, overseeing or designing/engineering autonomous systems are almost totally overcome but, the corollary of this is that, the advancement of digital warfare with high levels of artificially intelligent autonomous systems cannot surpass point that has already been reached. That is, the worry is not that we have reached another morally significant point in the history of weapons development, but that we are looking at a new enactment of an old trend. This is to say that to truly understand the relationship between soldiers, technical actors, autonomy and their ethicality for today’s purposes, we have to think about autonomous systems in the wider context of weapons that increase physical distance to the target and lower resistance to killing, as well how this maps onto increasing levels of artificial intelligence.

#### Autonomous weapons makes killing easy- diminishes humans to dots on a screen

Galliott 21 Galliott, Jai. "Humans, autonomous systems, and killing in war." Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics. IGI Global, 2021. 240-257. [AJL]

In order to fully understand the issues associated with autonomy, one must place lethal autonomous weapon systems on a spectrum with other forms of autonomous weaponry. At one end we have close range, then middle range and long range at other end. Close range, for our sake, involves any easily attributable kill at ‘point-blank’ range, whether with one’s bare hands, an edged weapon or even a projectile weapon. According to Grossman (1995), the key factor in close range killing is the undeniable responsibility one holds for the act. John Keegan and Richard Holmes (1986) cite the story of an Israeli paratrooper during the capture of Jerusalem in 1967: ‘we looked up at each other for half a second and I knew that it was up to me, personally, to kill him’ (p. 266). When a soldier kills at this range, more than any other, it’s an intensely vivid and personal matter (Grossman 1995, p. 115). One can see the raw emotions on their enemy’s face, hear their cries and smell the gunpowder. The Israeli paratrooper mentioned above goes on to say that having shot his enemy at close range, he could see the hate on his enemy’s face and that ‘there was so much blood...[he] vomited, until the rest of the boys came up’ (Keegan and Holmes 1986, p. 266). Combat at close proximity is an interpersonal affair, so much so that it is incredibly difficult to deny the enemy’s humanity. For this reason, Grossman says, the resistance to close-range killing is ‘tremendous’ (Grossman 1995, p. 118). At midrange – where you can still see and engage the enemy with handgrenades, sniper rifles and so on, but usually without being able to gauge the extent of the +XPDQV$XWRQRPRXV6\VWHPVDQG.LOOLQJLQ:DU wounds inflicted – the experience of killing changes. At this range in the spectrum, the soldier can begin to deny responsibility for the fatal shot or blow if there are others present and participating in the act of killing. One is still located on the battlefield and can hear the gunfire and feel the stress, but the distance between adversaries makes the act of killing both physically and psychologically easier, and thus more morally troubling (Grossman 1995, p. 113). At long range, at which one must use some sort of mechanical, electrical or digital assistance to view or interact with others and potential victims, there is evidence to suggest that killing is made even easier. Among those who we have historically considered least reluctant to kill are pilots, artillery numbers and missile silo attendees. Gwynne Dyer (2010, p. 57) writes that while being observed by their fellows puts pressure on them to kill (as was the case with the gunners in Vietnam), it has much more to do with the distance between them and their targets and how it acts as an emotional and moral buffer. She aptly notes that on the whole, ‘gunners fire at grid references they cannot see; submarine crews fire torpedoes at ‘ships’ (and not, somehow, at the people in the ships); and pilots launch their missiles at ‘targets’’ (Dywer 2010). Grossman (1995) also reports that in his extensive career researching and reading on the subject of killing in combat, he is not aware of a single instance in which an individual operating at such long range has refused to kill. We also have numerous examples of long distance killing made easy. Dyer (2010) reminds us that in the early nineteen forties, for instance, the British Royal Airforce ‘firebombed’ Hamburg. Using early bomber aircraft, munitions blew in windows and doors over four square miles and resulted in a firestorm which left seventy thousand people dead, mostly women, children and the elderly (Grossman 1995). A further eighty thousand died in the firebombing of Dresden, two hundred and twenty-five thousand in Tokyo and many millions more in bombing conflicts since (Grossman 1995). If the bomber crews had to kill each of these people with a flamethrower or, as Whetham (2012) writes, slit each of their throats with a knife, the majority would be unable to do it. The awfulness of killing people at such close proximity and the emotional trauma inherent to each act, and to the collective acts, would have been of such magnitude that they simply would not have happened. Figure 1. Distance, Technology and Resistance to Killing As indicated on the chart above, killing conducted with the most basic semi-autonomous systems belongs at the beginning end of the long-range killing spectrum, followed by higher level autonomous systems, which when roughly grouped together, might even be worthy of its own designation: killing at keyboard range. The contention here is that there is no other tactical weapon on the battlefield today that facilitates killing with such physical and psychological ease and that it becomes a rather clinical and dispassionate matter, easing any of their operators’ existing moral qualms. Noel Sharkey (2010) offers support this argument in drawing attention to reports collected by Singer. Amongst a variety of other disturbing statements, he cites one twenty-one year old soldier who talks about his acts of killing with casual indifference: ‘the truth is, it wasn’t all I thought it was cracked up to be. I mean, I thought killing somebody would be this life-changing experience. And then I did it, and I was like, ‘All right, whatever’ (Singer 2009, p. 391). Later, he says that ‘killing people is like squashing an ant. I mean, you kill somebody and it’s like ‘all right, let’s go get some pizza’ (Singer 2009, p. 392). In this clinical killing environment, in which it seems reasonable to propose that some human targets are divested of their humanity, respect for the laws of war wanes. Many ‘war porn’ videos show raw footage of Predator strikes with people being reduced to little more than ‘hot spots’ or ‘blips’ on the screen, with operators often failing to take the necessary precautions to ensure noncombatants are protected. And, of course, it is not difficult to imagine how someone coding an artificially intelligence autonomous weapon might come to minimise the impact of their effort in reducing decisions regarding people to keystrokes. Again, it seems that obscuring ‘targets’ in this way and increasing the distance to the maximum possible, makes it easier for to kill in an indiscriminate and disproportionate fashion.

#### Autonomous weapons exacerbates racism in war- perpetuates idea of “the other”

Galliott 21 Galliott, Jai. "Humans, autonomous systems, and killing in war." Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics. IGI Global, 2021. 240-257. [AJL]

While the distances involved show the powerful role of distance and autonomous systems in overcoming moral-emotional qualms and the socialised inhibition to killing, there are a range of other mechanisms that further this and make it even easier for systems operators, designers, programmers and the like to contribute to deployment of lethal force without regard for the consequences of their individual actions or jus in bello norms. The first additional mechanism that Grossman (1995, p. 161) explicitly notes – and the one that is most relevant to America’s use of autonomous platforms, but will only mention briefly here – is cultural difference, which perpetuates racial and ethnic differences and allows warfighters to further dehumanise the enemy. Military forces have long been trying to get their troops to think of the enemy/ies as ‘inferior forms of life...as less than human’ (Watson 1978, p. 250). Put simply, the further one is technologically distanced from the enemy, the easier it is to think that they are distinctly different to you in some way and the easier it is to kill them. Unmanned systems separate soldiers from the cultural environment that they would operate in if they were in the field and, in that sense, permit them to racially demonise and ‘other’ the enemy. Connected to cultural distance is moral distance, which involves legitimising oneself and one’s cause (Grossman 1995, p. 164). Once it has been determined that the enemy is culturally inferior, it is not difficult for operators of unmanned platforms to incorrectly suppose that their counterparts are either misguided or share their leaders’ moral guilt and think that this warrants waging unconstrained violence against those with the supposedly morally inferior cause (Grossman 1995, 164-7). In the case of highly autonomous systems, it may involve a less overt process, whereby contributing technical actors find it difficult to account for cultural differences in code, whether it be in a sensing or action mode, resulting in the same consequences.

#### LAWs turn case – it leads to more conflict since nations will view war as a more viable option and risks falling into the wrong hands risking terror attacks and even genocide

**Anzarouth 21** (Matthew Anzarouth, 12-2-2021, accessed on 6-25-2022, Harvard Political Review, "Robots that Kill: The Case for Banning Lethal Autonomous Weapon Systems - Harvard Political Review", https://harvardpolitics.com/robots-that-kill-the-case-for-banning-lethal-autonomous-weapon-systems/)-qcl

In the days leading up to its withdrawal from Afghanistan, the U.S. military conducted a drone strike that killed 10 civilians in Kabul. The timing of this tragedy, in the midst of the mass evacuation from Afghanistan, casts doubt on the U.S. military’s promise to stop serving as a global policeman. The Biden administration has not ended the “forever wars” — it has simply elected to fight them with robots in the sky rather than boots on the ground. Pointing to the drone strike in Kabul as prime evidence, many experts warn of the dangers of Biden’s ‘over-the-horizon’ counterterrorism strategy, which uses imprecise semi-autonomous drones to replace human soldiers and combat terrorists from afar. Little attention, however, is being paid to an even more threatening weapon that may define the coming decades of war. Soon, guided missiles and semi-autonomous drones may be replaced by fully autonomous weapons that have the ultimate say over who lives and who dies. Lethal autonomous weapon systems are being introduced into military arsenals, and the United States, Russia, South Korea, Israel and the United Kingdom have shown a keen interest in their development. Unlike semi-autonomous drones, LAWS can select targets and attack them without any human intervention. These weapons are still in their infancy and, over time, will likely develop greater autonomy and more capabilities. One type of autonomous weapon would, after being activated by a human operator, fly around the world, identify its targets and fire missiles at them. An existing preliminary version of this weapon is the Israeli Harpy, which is programmed to roam around in the air in a predetermined loitering area, detecting and attacking enemy radar emitters. Political scientist Michael C. Horowitz posits in Dædalus that as technology progresses militaries may even use LAWS that serve as operations planning systems, autonomous battle systems that “could decide the probability of winning a war and whether to attack, plan an operation and then direct other systems — whether human or robotic — to engage in particular attacks.” The appeal of LAWS to countries like the U.S. and Russia is quite intuitive. If a country can fight wars with ruthless efficiency, accurately pick out terrorists from hundreds of feet in the sky, and spare the lives of thousands of soldiers, why wouldn’t it do so? A closer inspection reveals that the costs of this technology vastly outweigh the benefits. The Danger in Killer Robots The use of LAWS would lower the threshold for states going to war, increasing the likelihood of conflict. Many philosophers, political scientists and governments have expressed the concern that militaries will resort to conflict more often if they do not need to rely on soldiers and can use LAWS instead. Domestic populations will be less wary of conflict if it no longer means seeing fellow citizens risk their lives on the battlefield. The threshold-lowering effect of LAWS is particularly relevant in the context of a current bipartisan trend in the U.S. against intervention. It is plausible that without LAWS, the era of U.S. unilateral interventions and the war on terror would come to an end. Recognizing the failures of wars in Vietnam, Iraq and Afghanistan, politicians on both sides of the political spectrum are pushing not to send troops abroad to risk their lives. But the option of using LAWS and sidestepping the costs to a country’s soldiers threatens to reverse this anti-war trend and provide militaries with a politically palatable way of fighting wars. There could be catastrophic consequences if we liberate militaries from political constraints preventing them from going to war. The first wave of the proliferation of LAWS may simply look like the natural progression of our current drone capabilities. For instance, Russia may have already used autonomous drones to attack targets in Syria, but these weapons are only different from current semi-autonomous drones in the greater degree of risk assumed by eliminating human intervention. In other instances, however, the use of LAWS will present substantial advantages that make them different in kind from drones as we know them. Consider, for example, Azerbaijan’s use of Israeli-supplied IAI Harop drones in the war with Armenia in 2020. The loitering munition system used by the military allowed tiny and hardly-detectable autonomous drones to circle over the enemy’s defense line, pick out targets and attack them, an ability that proved decisive in Azerbaijan’s victory in the war. To understand what a world with LAWS will look like in the long term requires a bit of imagination. Perhaps a post-withdrawal Afghanistan will involve weapons like the Harop drones constantly roaming the skies and diving into the ground to take out targets. Or maybe we will see the chilling predictions of science fiction come true. In their book AI 2041, writers Chen Qiufan and Kai-Fu Lee express their fear that LAWS will fall into the hands of armed groups and terrorists. They describe a “Unabomber-like scenario in which a terrorist carries out the targeted killing of business elites and high-profile individuals,” using autonomous drones that rely on facial recognition to identify their targets. Leading expert in artificial intelligence Toby Walsh warns of these weapons falling into the hands of dictators and being used as tools of ethnic cleansing. Even if we assume that LAWS are operated primarily by legitimate militaries, additional complications arise when we consider what happens in the case of unjust killings. Philosopher Robert Sparrow argues that the autonomy of LAWS makes it impossible to hold anyone accountable for illegitimate killings they commit. If the robot acted autonomously, tracing accountability back to another agent seems morally objectionable and legally infeasible. But it would also be unjust to not punish illegitimate killings. This dilemma presents a so-called ‘responsibility gap’, where no one can be held responsible for illegitimate killings, and wrongful acts of war go undeterred. Preventing The Next Arms Race Despite these grave concerns, countries are pushing ahead in the research and development of LAWS. With large military powers leading the race, there are two potential outcomes if this trend goes uninterrupted. One is that LAWS become tools with which powerful militaries destabilize other regions, starting a new chapter of the ‘forever wars’ without boots on the ground. The second potential outcome is that LAWS become front and centre in conflict between the large military powers leading the race. They may drag us into a new war between superpowers without the mutually assured destruction that prevents nuclear warfare since LAWS can engage in a series of smaller, yet still extremely impactful, attacks that will not be deterred by the threat of retaliation.

### FRT

#### FRT perpetuates racism and misogyny through inherent bias

**Wolfe & Dastin 19** (Jan Wolfe Reuters Jounralist, Jeffrey Dastin, Technology Correspondent at Reuters News Agency 12-20-2019, accessed on 6-25-2022, Reuters, "U.S. government study finds racial bias in facial recognition tools", https://www.reuters.com/article/us-usa-crime-face/u-s-government-study-finds-racial-bias-in-facial-recognition-tools-idUSKBN1YN2V1)-qcl

The study by the National Institute of Standards and Technology (NIST) found that, when conducting a particular type of database search known as “one-to-one” matching, many facial recognition algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces. The study also found that African-American females are more likely to be misidentified in “one-to-many” matching, which can be used for identification of a person of interest in a criminal investigation. While some companies have played down earlier findings of bias in technology that can guess an individual’s gender, known as “facial analysis,” the NIST study was evidence that face matching struggled across demographics, too. Joy Buolamwini, founder of the Algorithmic Justice League, called the report “a comprehensive rebuttal” of those saying artificial intelligence (AI) bias was no longer an issue. The study comes at a time of growing discontent over the technology in the United States, with critics warning it can lead to unjust harassment or arrests. For the report, NIST tested 189 algorithms from 99 developers, excluding companies such as Amazon.com Inc [AMZN.O](https://www.reuters.com/companies/AMZN.O) that did not submit one for review. What it tested differs from what companies sell, in that NIST studied algorithms detached from the cloud and proprietary training data. China’s SenseTime, an AI startup valued at more than $7.5 billion, had “high false match rates for all comparisons” in one of the NIST tests, the report said. SenseTime’s algorithm produced a false positive more than 10% of the time when looking at photos of Somali men, which, if deployed at an airport, would mean a Somali man could pass a customs check one in every 10 times he used passports of other Somali men. SenseTime said the report “reflects an isolated case” and that what it submitted had bugs it has addressed. “The results are not reflective of our products, as they undergo thorough testing before entering the market (and) all report a high degree of accuracy,” it said. Yitu, another AI startup from China, was more accurate and had little racial skew. Microsoft Corp [MSFT.O](https://www.reuters.com/companies/MSFT.O) had almost 10 times more false positives for women of color than men of color in some instances during a one-to-many test. Its algorithm showed little discrepancy in a one-to-many test with photos just of black and white males. Microsoft said it was reviewing the report and did not have a comment on Friday morning. Congressman Bennie Thompson, chairman of the U.S. House Committee on Homeland Security, said the findings of bias were worse than feared, at a time when customs officials are adding facial recognition to travel checkpoints. “The administration must reassess its plans for facial recognition technology in light of these shocking results,” he said.

### Surveillance

#### Surveillance capitalism weaponizes digital technology to control communities and turns democracy promotion – US/Big tech partnership reproduces economic inequality, human rights violation and environmental degradation

Hynes 21

[Mike](file:////insight/search%3fq=Mike%20Hynes) Hynes, PhD, is  lecturer in Political Science and Sociology at the National University of Ireland Galway, specializing in environmental sociology, Mobilities and sustainability research. (2021). "Digital Democracy: The Winners and Losers", The Social, Cultural and Environmental Costs of Hyper-Connectivity: Sleeping Through the Revolution, Emerald Publishing Limited, Bingley, pp. 137-153. <https://doi.org/10.1108/978-1-83909-976-220211009> --JrH

The digital ICT revolution promised much for democratic politics in the twenty-first century but so far has delivered little but disruption. The dawn of the internet age was to bring a decisive shift towards the citizen and information was to become free and limitless, and enlightenment and empowerment would follow. But while digital technologies provide us with the opportunity to accumulate quantities of information that one time may not have been possible, big tech and the state remains much better equipped than any private citizen to take full advantage of this opportunity. In many ways, digital technology has been weaponised against the very system it was purported to support and defend and the citizens it was meant to engage, protect and enlighten. Authoritarian regimes across the world have seized upon the opportunities provided by such technology to increase surveillance and control of their people while simultaneously spreading misinformation and confusion, undermining many of the established Western liberal democracies. It would be rather naïve to think that democratic governments are not also regularly using similar digital surveillance technique under various guises and security apparatuses. And all the while big tech is the real big winner. The pioneers of surveillance capitalism Google were emboldened and benefitted from historical events when a national security apparatus, galvanised by the attacks of 11 September 2001, saw the emergent capabilities and the promise of some certainty in how Google’s storage and use of huge stocks of personal data could be used to shadow and predict the behaviour of individuals.[37](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn37)Zuboff believes that the concepts underpinning surveillance capitalism are facilitating the overthrow of the people’s sovereignty and is a prominent force in the perilous drift towards democratic deconsolidation that now threatens Western liberal democracies themselves. And this is a common complaint in the twenty-first century; democracy itself has lost control of corporate power in the form of big tech companies, who use whatever means possible to hoard vast wealth and influence while fuelling inequality, damaging the planet and avoid paying their fair share of taxes.[38](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn36)Today’s big tech behemoths exist in a political culture that has grown accustomed and accommodating to their every need, and Runciman argues, in the United States, this was further cemented by the Supreme Court decision in the Citizens United case of 2010 to grant corporations the same rights to free speech as individual citizens.[39](https://www.emerald.com/insight/content/doi/10.1108/978-1-83909-976-220211009/full/html#fn39) The ideals and very notion of liberal democracy are now under constant pressure from many angles, and the traditional hierarchy of power is also under increasing danger. The power of modern corporate power, in the form of big tech, has grown exponentially over the past decade to the point where it now has the wherewithal to undermine how democracy itself operates and not be overly worried about the consequences. A major imperative now for every citizen and democratic nation must be to reassess the inequitable influence of big tech corporate power and the internet, particularly as it relates to our personal data, and to question: who owns and controls such power, and what right do they have to use and misuse our personal data to undermine our key democratic institutions? Democracy must be seen to represent the wishes of the people rather than viewed as a system of corporate tyranny.

#### “Digital democracy” is a myth – US mass surveillance, weakening data privacy protections, banning location all expose the coercive hypocrisy of US “digital authoritarianism”

Klyman 22

Kevin Klyman researches US-China relations and has written data protection policies adopted by the World Health Organization. “Biden’s Campaign for “Digital Democracy” Is Really a Giveaway to Big Tech”, Jacobin. 6/26/22. <https://jacobin.com/2022/06/us-tech-companies-government-contracts-data-google-facebook-microsoft-amazon/> --JrH

Mejias asked Jacobin, “How can ‘digital democracy’ be so different from ‘digital authoritarianism’ if they run practically on the same technologies and infrastructures? We should be concerned about how easy it would be, technologically, to switch from the former to the latter.” “Belarus used equipment provided by the American company Sandvine to block millions of websites and censor news and social media in the midst of mass protests.” In the United States, algorithmic decision making is already used to [incarcerate](https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html) people, [deny](https://www.wsj.com/articles/robots-are-taking-over-the-rental-screening-process-11574332200) them housing, and [determine](https://www.wsj.com/articles/how-hospitals-are-using-ai-to-save-lives-11649610000) if they receive medical care. The US government’s surveillance capabilities are only growing. The NSA’s surveillance programs are [entirely intact](https://www.washingtonpost.com/national-security/nsa-surveillance-xkeyscore-privacy/2021/06/29/b2134e7a-d685-11eb-a53a-3b5450fdca7a_story.html)nine years after Edward Snowden exposed them, allowing the US government to covertly [collect](https://theintercept.com/2015/07/01/nsas-google-worlds-private-communications/) the rest of the world’s internet traffic, read people’s emails, and hack into their devices. In addition, key provisions of the Patriot Act — which [empowers](https://www.aclu.org/other/surveillance-under-usapatriot-act) the government to spy on Americans and people outside the United States with no oversight — [remain](https://www.salon.com/2021/10/16/after-20-years-its-time-to-repeal-the-patriot-act-and-begin-to-dismantle-the-surveillance-state/) in effect. US tech companies devour as much sensitive data as possible. The [vast majority](https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/digital%20globalization%20the%20new%20era%20of%20global%20flows/mgi-digital-globalization-full-report.ashx) of all cross-border data traffic passes through the United States, with 70 percent of global internet traffic [passing](https://www.theatlantic.com/technology/archive/2016/01/amazon-web-services-data-center/423147/) through northern Virginia, the headquarters of many US intelligence agencies. Microsoft, Amazon, and Google [own](https://www.srgresearch.com/articles/microsoft-amazon-and-google-account-for-over-half-of-todays-600-hyperscale-data-centers) more than half of the world’s major data centers, and along with Meta, they [own](https://www.wsj.com/articles/google-amazon-meta-and-microsoft-weave-a-fiber-optic-web-of-power-11642222824) the rights to two-thirds of the capacity of the world’s undersea fiber-optic cables, which form the [backbone](https://restofworld.org/2022/google-meta-underwater-cables/) of the internet. As Google’s cofounder Larry Page [predicted](https://www.nytimes.com/2021/11/12/opinion/facebook-privacy.html) twenty years ago, “Everything you’ve ever heard or seen or experienced will become searchable. Your whole life will be searchable.” Government of the People, by the Algorithms, for the Corporations America’s technology policy priorities abroad encompass three key objectives: banning data localization, weakening data protection laws, and eliminating digital services taxes. This deregulatory agenda would eliminate nontariff barriers employed by poor countries to develop their local tech ecosystems and keep predatory multinationals at bay. The US government and Big Tech have relentlessly pressured countries to abstain from adopting such measures — purportedly in the interest of democracy. Data localization regulations [require](https://www.cloudflare.com/learning/privacy/what-is-data-localization/) that data generated in a country is stored or processed in that country. The Center for Strategic and International Studies — which Facebook [funds](https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/201015_Yayboke_Brannen_PromoteAndBuild_Brief.pdf) to work on data localization and digital authoritarianism — [argues](https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/210723_Sheppard_DataLocalization.pdf?en2io56tR_AVK4Ts6yzoHoafKr354j5t) that “data localization can be used as a tool of digital authoritarianism to limit democracy and human rights. . . . This is intended to facilitate these governments’ ability to carry out ‘a crackdown on free expression, privacy, and a range of human rights.’” This is undoubtedly true in many cases. China’s 2017 cybersecurity law [requires](https://www.hrw.org/news/2016/11/06/china-abusive-cybersecurity-law-set-be-passed) that foreign firms not only store data within China but also provide technical support to Chinese security services conducting investigations. But data localization can also have significant economic benefits for low-income countries. Requirements to store data where it is collected force multinationals to invest in local data centers and data processors, often creating more competition among local companies that store data. Local data storage also [increases](https://www.ft.com/content/adb1130e-2844-4051-b1df-a691fc8a19b8) the speed of digital services since the information that fuels applications does not have to be fetched from halfway across the world. Moreover, it lowers [costs](https://www.internetsociety.org/wp-content/uploads/2020/06/Anchoring-the-African-Internet-Ecosystem-Lessons-from-Kenya-and-Nigeria.pdf) because data does not have to pass through expensive international transit points. This is, in part, why the European Union has adopted de facto data localization requirements. In 2020, the total capacity of data centers in sub-Saharan Africa was just one-quarter that of the city of London. To rectify this gap, Africa’s largest economies, such as [South Africa](https://itif.org/publications/2021/07/19/how-barriers-cross-border-data-flows-are-spreading-globally-what-they-cost), [Nigeria](https://www.uubo.org/media/2296/data-localization-laws-nigeria.pdf), [Kenya](https://www.ids.ac.uk/opinions/lobbying-for-digital-dominance-in-africa/), Algeria, and [Rwanda](https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2019/03/GSMA_Understanding-the-impact-of-data-localisation.pdf), have adopted data localization rules, while countries like Ghana are in the process of implementing similar measures. These laws have [fueled](https://www.ft.com/content/402a18c8-5a32-11ea-abe5-8e03987b7b20) a gold rush on the continent, bringing in billions of dollars of investment and [doubling](https://www.economist.com/middle-east-and-africa/2021/12/04/data-centres-are-taking-root-in-africa) local data center capacity in the last five years. In Kenya and Nigeria, the price of data has [fallen](https://www.internetsociety.org/wp-content/uploads/2020/06/Anchoring-the-African-Internet-Ecosystem-Lessons-from-Kenya-and-Nigeria.pdf) by 90 percent, due in large part to localization.Data localization can also deliver myriad noneconomic benefits. Storing data locally makes it [more difficult](https://www.globaljustice.org.uk/sites/default/files/files/resources/e-pocalypse_now_briefing.pdf) for foreign governments to surveil a country’s citizens because they must do so from a significantly greater distance. In the [case](https://academic.oup.com/ijlit/article-abstract/25/3/213/3960261?redirectedFrom=PDF) of the United States, “It is much cheaper for the FBI to issue a national security letter compelling a US-based data host to provide access to their data centres than it is for the NSA to gain access to data stores outside of the USA.” Furthermore, data that is stored locally can be audited by local regulators, enabling them to carry out their democratically mandated functions. Nick Dearden, the director of Global Justice Now, told Jacobinthat measures banning data localization are “going to make it impossible for regulators to properly control what Big Tech is doing and constrain these companies’ activities in the public interest.” Free-market ideologues have dismissed these justifications for data localization measures, paternalistically [claiming](https://www2.itif.org/2017-cross-border-data-flows.pdf?_ga=2.243851781.1067956188.1647786604-1495566515.1646079866) that poor countries are “shooting themselves in the foot” by pursuing “digital mercantilism.” The Information Technology and Innovation Foundation has even gone so far as to [argue](https://www2.itif.org/2013-localization-barriers-to-trade.pdf) that the World Bank, International Monetary Fund, OECD, and US Agency for International Development “all need to cut off foreign aid” to countries with data localization laws.

#### AI used unethically to gather data on citizens

Schneier 16 Bruce Schneier, 02-08-2016, " Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World" W. W. Norton & Company, <https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath_-2015.pdf> [AJL]

“Governments and corporations gather, store, and analyze the tremendous amount of data we chuff out as we move through our digitized lives. Often this is without our knowledge, and typically without our consent. Based on this data, they draw conclusions about us that we might disagree with or object to, and that can impact our lives in profound ways. We may not like to admit it, but we are under mass surveillance. Much of what we know about the NSA’s surveillance comes from Edward Snowden, although people both before and after him also leaked agency secrets. As an NSA contractor, Snowden collected tens of thousands of documents describing many of the NSA’s surveillance activities. In 2013, he fled to Hong Kong and gave them to select reporters. For a while I worked with Glenn Greenwald and the Guardian newspaper, helping analyze some of the more technical documents. The first news story to break that was based on the Snowden documents described how the NSA collects the cell phone call records of every American. One government defense, and a sound bite repeated ever since, is that the data collected is “only metadata.” The intended point was that the NSA wasn’t “collecting the words we spoke during our phone conversations, only the phone numbers of the two parties, and the date, time, and duration of the call. This seemed to mollify many people, but it shouldn’t have. Collecting metadata on people means putting them under surveillance. An easy thought experiment demonstrates this. Imagine that you hired a private detective to eavesdrop on someone. The detective would plant bugs in that person’s home, office, and car. He would eavesdrop on that person’s phone and computer. And you would get a report detailing that person’s conversations. Now imagine that you asked the detective to put that person under surveillance. You would get a different but nevertheless comprehensive report: where he went, what he did, who he spoke with and for how long, who he wrote to, what he read, and what he purchased. That’s metadata. Eavesdropping gets you the conversations; surveillance gets you everything else. Telephone metadata alone reveals a lot about us. The timing, length, and frequency of our conversations reveal our relationships with others: our intimate friends, business associates, and everyone in-between. Phone metadata reveals what and who we’re interested in and what’s important to us, no matter how “private. It provides a window into our personalities. It yields a detailed summary of what’s happening to us at any point in time.”

#### AI systems eliminate the possibility of consent based on the sheer amount of data collected and utilized for services

Varon & Peña 21 (Joana, Executive Directress and Creative Chaos Catalyst at Coding Right, Paz Latin American Institute of Terraforming, “Artificial intelligence and consent: A feminist anticolonial critique”, Econstor, Institute for Internet and Society, Berlin, Vol. 10, Iss. 4, https://doi.org/10.14763/2021.4.1602 , pp. 8-11)-qcl

Traditionally, it has been considered by data protection regulations that there is an invasion of privacy if there is no consent from the data owner to the data processor unless there are legal obligations, vital interests, public interest, or legitimate interests. These are also some of the legal bases for processing personal data under several acts of data protection legislation compatible with the General Data Protection Regulation (GDPR). While being presented as the primary basis for data processing, meaningful consent in the use of personal data in digital services has been largely problematised as ineffective (Lee et al, 2017). But the already known problems such as notification, choice, and proper withdrawal of consent (Jones et al., 2018) can be exacerbated by artificial intelligence systems that collect huge amounts of data, process and generate new data. In this context, even if AI system controllers really want to obtain transparent and meaningful consent, they just cannot do it because they don’t know where data is going and how it’s going to be utilised (Nissenbaum, 2018). Furthermore, controllers of these systems also say they don’t have the ability to inform us about the risks we are consenting to, not necessarily as a matter of bad faith, but because of increasingly powerful computational methods such as machine learning working as a black box (Tufekci, 2018; Carmi, 2020). For other authors, the unpredictable and even unimaginable use of data by AI systems are even considered a feature, not a bug. For this same reason, companies and parties collecting and processing data have an incentive to leave unspecified the range of potential future applications (Jones et al., 2018; Cohen, 2018). This system’s opacity has been considered a major problem for meaningful consent, for example, regarding the uses of AI in medical diagnosis consultations (Astromskė et al., 2020). Even so, the criticism of consent in AI is still not very extensive, and it is largely influenced by the criticism of digital consent, focusing on the transparency and unpredictability aspects of the systems. Much of the concerns around consent on data processing have been approached by self-regulation solutions, the Federal Trade Commission in the United States being one of its main sponsors. For researcher Daniel Solove (2013), under the current approach of privacy regulation—that he would call “privacy self-management”, but is also called “privacy as control” by other scholars (Cohen, 2018)—policymakers try to provide people with a set of rights to enable them to make decisions about how to manage their data. This is an individual framing of consent, based on the assumption that we are all autonomous, free, and rational individuals with the capacity to consent, disregarding our possibility of doing so due to unequal power dynamics. Two have been the main measures of mitigation in this framework of self-regulation: anonymisation and transparency and choice (also called notice & consent) (Barocas & Nissenbaum, 2009; Nissenbaum, 2011). For Barocas and Nissenbaum (2009), this approach has an appeal to stakeholders and regulators basically because notice and consent—as a way to give individual control to users—seems to adequately fit in the popular definition of privacy as a right to control information about oneself. In the same way, notice and consent seem consistent with the idea of a free market, “because personal information may be conceived as part of the price of online exchange, all is deemed well if buyers are informed of a seller’s practices collecting and using personal information and are allowed freely to decide if the price is right” (Nissenbaum, 2011, p. 34). In general terms, the critical voices on the model of notice and consent could be divided into two general groups: One that we call—borrowing the denomination from Nissenbaum (2011)—“critical adherents”, which are moderate in their critics and focus on improving procedures of the model of consent, more than criticizing the liberal paradigm. While the other group is much more radical in terms of not believing at all in the model of notice and consent, basically because they don't believe in the paradigm of privacy as individual control and autonomy. The main criticism of critical adherents focuses on the way consent is being offered to citizens. They are critical about the idea of consent as “take it or leave it” and believe in a more granular model of consent (Solove, 2013). They are also critical about the idea of choice as “opt-out” and push for a model of “opt-in” (Nissenbaum, 2011; Hotaling, 2008). Likewise, this group acknowledges that privacy policies are long, legalistic, and really hard to digest for regular citizens; it is also an unrealistic burden for individuals to notice and review hundreds of online contracts from start to finish (Hotaling, 2008) and, in this context, they also advocate increasing transparency (Nissenbaum, 2011). Nevertheless, in addition to its unpredictability and opacity, artificial intelligence brings new challenges to the classic free model of notification and consent. AI systems applied to social programmes can induce personal information from individuals in unexpected and even manipulative ways. And also, many of these applications challenge the form of screen-based notification and consent model, since, most of the time, it is not a software that has direct interaction with the users who feed the system with their data, for instance, when they rely on technologies such as facial recognition or the “Internet of Things” (Jones at al., 2018). For more severe critics of liberal consent, meaningful consent requires meaningful notice. In reality, the information provided about data collection, its processing, and use tends to be vague and general, or too cryptic for non-lawyers. For Nissenbaum (2011), the traditional notion behind “online privacy” suggests that “online” is a distinctive sphere where protecting personal information is always framed in the context of commercial online transactions. As we have mentioned before, Julie E. Cohen goes further and considers privacy as an environmental condition (Cohen, 2012, 2018). Thus, protecting privacy effectively requires a willingness to depart more definitively from subject-centred frameworks in favour of condition-centred frameworks (Cohen, 2018). Therefore, only this form of criticism considers structural power relations when addressing consent and data processing. Following Cohen, Carmi (2018) goes even further and stresses that meanwhile legal and tech narratives frame online consent as if people—their data self or data bodies—were a defined, static, and almost tangible piece of personal property, our everyday realities as subjects are far away from that: we present ourselves in a fluid—never fixed—way depending on the context. Static categorisation, hierarchical evaluation according to the values of those in power and separation of different human beings to be targeted for surveillance and control was at the heart of colonisation practices. It is again at the heart of Digital Welfare States, as this is exactly what predictive algorithms and risk modeling systems operated by these welfare programmes are doing to determine social services, affecting a wide variety of aspects in life: work conditions, pensions, education, health, support for people with disability, and many others.

#### AI leads to violence towards innocent civilians

Schneier 16 Bruce Schneier, 02-08-2016, " Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World" W. W. Norton & Company, [https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath\_-2015.pdf pg 69-70](https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath_-2015.pdf%20pg%2069-70). [AJL]

Another harm of government surveillance is the way it leads to people’s being categorized and discriminated against. George Washington University law professor Daniel Solove calls the situation Kafkaesque. So much of this data is collected and used in secret, and we have no right to refute or even see the evidence against us. This will intensify as systems start using surveillance data to make decisions automatically. Surveillance data has been used to justify numerous penalties, from subjecting people to more intensive airport security to deporting them. In 2012, before his Los Angeles vacation, 26-year-old Irishman Leigh Van Bryan tweeted, “Free this week, for quick gossip/prep before I go and destroy America.” The US government had been surveilling the entire Twitter feed. Agents picked up Bryan’s message, correlated it with airplane passenger lists, and were waiting for him at the border when he arrived from Ireland. His comment wasn’t serious, but he was questioned for five hours and then sent back home. We know that bomb jokes in airports can get you detained; now it seems that you have to be careful making even vague promises of international rowdiness anywhere on the Internet. In 2013, a Hawaiian man posted a video on Facebook showing himself drinking and driving. Police arrested him for the crime; his defense was that it was a parody and that no actual alcohol was consumed on the video. It’s worse in the UK. There, people have been jailed because of a racist tweet or a tasteless Facebook post. And it’s even more extreme in other countries, of course, where people are routinely arrested and tortured for things they’ve written online. Most alarming of all, the US military targets drone strikes partly based on their targets’ data. There are two types of drone targeting. The first is “targeted killing,” where a known individual is located by means of electronic or other surveillance. The second is “signature strikes,” where unidentified individuals are targeted on the basis of their behavior and personal characteristics: their apparent ages and genders, their location, what they appear to be doing. At the peak of drone operations in Pakistan in 2009 and 2010, half of all kills were signature strikes. We don’t have any information about how accurate the profiling was. This is wrong. We should be free to talk with our friends, or send a text message to a family member, or read a book or article, without having to worry about how it would look to the government: our government today, our government in five or ten years, or some other government. We shouldn’t have to worry about how our actions might be interpreted or misinterpreted, or how they could be used against us. We should not be subject to surveillance that is essentially indefinite.

#### AI used to unethically monitor innocent civilians

Schneier 16 Bruce Schneier, 02-08-2016, " Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World" W. W. Norton & Company, [https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath\_-2015.pdf pg -77-780](https://ciberativismoeguerra.files.wordpress.com/2017/09/bruce-schneier-data-and-goliath_-2015.pdf%20pg%20-77-780)[AJL]

Aside from such obvious abuses of power, there’s the inevitable expansion of power that accompanies the expansion of any large and powerful bureaucratic system: mission creep. For example, after 9/11, the CIA and the Treasury Department joined forces to gather data on Americans’ financial transactions, with the idea that they could detect the funding of future terrorist groups. This turned out to be a dead end, but the expanded surveillance netted a few money launderers. So it continues. In the US, surveillance is being used more often, in more cases, against more offenses, than ever before. Surveillance powers justified in the PATRIOT Act as being essential in the fight against terrorism, like “sneak and peek” search warrants, are far more commonly used in non-terrorism investigations, such as searches for drugs. In 2011, the NSA was given authority to conduct surveillance against drug smugglers in addition to its traditional national security concerns. DEA staff were instructed to lie in court to conceal that the NSA passed data to the agency. The NSA’s term is “parallel construction.” The agency receiving the NSA information must invent some other way of getting at it, one that is admissible in court. The FBI probably got the evidence needed to arrest the hacker Ross Ulbricht, aka Dread Pirate Roberts, who ran the anonymous Silk Road website where people could buy drugs and more, in this way. Mission creep is also happening in the UK, where surveillance intended to nab terrorists is being used against political protesters, and in all sorts of minor criminal cases: against people who violate a smoking ban, falsify their address, and fail to clean up after their dogs. The country has a lot of cameras, so it “makes sense” to use them as much as possible. Other countries provide many more examples. Israel, for instance, gathers intelligence on innocent Palestinians for political persecution. Building the technical means for a surveillance state makes it easy for people and organizations to slip over the line into abuse. Of course, less savory governments abuse surveillance as a matter of course—with no legal protections for their citizens. All of this matters, even if you happen to trust the government currently in power. A system that is overwhelmingly powerful relies on everyone in power to act perfectly—so much has to go right to prevent meaningful abuse. There are always going to be bad apples—the question is how much harm they are allowed and empowered to do and how much they corrupt the rest of the barrel. Our controls need to work not only when the party we approve of leads the government but also when the party we disapprove of does.

## Impacts

### War

#### AI extremely unpredictable and dangerous – used to fuel endless war

Santa Clara University ND Santa Clara University, No date, "Artificial Intelligence and Ethics: Sixteen Challenges and Opportunities," No Publication, <https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics-sixteen-challenges-and-opportunities/> [AJL]

Artificial intelligence and machine learning technologies are rapidly transforming society and will continue to do so in the coming decades. This social transformation will have deep ethical impact, with these powerful new technologies both improving and disrupting human lives. AI, as the externalization of human intelligence, offers us in amplified form everything that humanity already is, both good and evil. Much is at stake. At this crossroads in history we should think very carefully about how to make this transition, or we risk empowering the grimmer side of our nature, rather than the brighter. Why is AI ethics becoming a problem now? Machine learning (ML) through neural networks is advancing rapidly for three reasons: 1) Huge increase in the size of data sets; 2) Huge increase in computing power; 3) Huge improvement in ML algorithms and more human talent to write them. All three of these trends are centralizing of power, and “With great power comes great responsibility” [2]. As an institution, the Markkula Center for Applied Ethics has been thinking deeply about the ethics of AI for several years. This article began as presentations delivered at academic conferences and has since expanded to an academic paper (links below) and most recently to a presentation of “Artificial Intelligence and Ethics: Sixteen Issues” I have given in the U.S. and internationally [3]. In that spirit, I offer this current list: 1. Technical Safety The first question for any technology is whether it works as intended. Will AI systems work as they are promised or will they fail? If and when they fail, what will be the results of those failures? And if we are dependent upon them, will we be able to survive without them? For example, several people have died in a semi-autonomous car accident because vehicles encountered situations in which they failed to make safe decisions. While writing very detailed contracts that limit liability might legally reduce a manufacturer’s responsibility, from a moral perspective, not only is responsibility still with the company, but the contract itself can be seen as an unethical scheme to avoid legitimate responsibility. The question of technical safety and failure is separate from the question of how a properly-functioning technology might be used for good or for evil (questions 3 and 4, below). This question is merely one of function, yet it is the foundation upon which all the rest of the analysis must build. 2. Transparency and Privacy Once we have determined that the technology functions adequately, can we actually understand how it works and properly gather data on its functioning? Ethical analysis always depends on getting the facts first—only then can evaluation begin. It turns out that with some machine learning techniques such as deep learning in neural networks it can be difficult or impossible to really understand why the machine is making the choices that it makes. In other cases, it might be that the machine can explain something, but the explanation is too complex for humans to understand. For example, in 2014 a computer proved a mathematical theorem, using a proof that was, at the time at least, longer than the entire Wikipedia encyclopedia [4]. Explanations of this sort might be true explanations, but humans will never know for sure. As an additional point, in general, the more powerful someone or something is, the more transparent it ought to be, while the weaker someone is, the more right to privacy he or she should have. Therefore the idea that powerful AIs might be intrinsically opaque is disconcerting. 3. Beneficial Use & Capacity for Good The main purpose of AI is, like every other technology, to help people lead longer, more flourishing, more fulfilling lives. This is good, and therefore insofar as AI helps people in these ways, we can be glad and appreciate the benefits it gives to us. Additional intelligence will likely provide improvements in nearly every field of human endeavor, including, for example, archaeology, biomedical research, communication, data analytics, education, energy efficiency, environmental protection, farming, finance, legal services, medical diagnostics, resource management, space exploration, transportation, waste management, and so on. As just one concrete example of a benefit from AI, some farm equipment now has computer systems capable of visually identifying weeds and spraying them with tiny targeted doses of herbicide. This not only protects the environment by reducing the use of chemicals on crops, but it also protects human health by reducing exposure to these chemicals. 4. Malicious Use & Capacity for Evil A perfectly well functioning technology, such as a nuclear weapon, can, when put to its intended use, cause immense evil. Artificial intelligence, like human intelligence, will be used maliciously, there is no doubt. For example, AI-powered surveillance is already widespread, in both appropriate contexts (e.g., airport-security cameras), perhaps inappropriate ones (e.g., products with always-on microphones in our homes), and conclusively inappropriate ones (e.g., products which help authoritarian regimes identify and oppress their citizens). Other nefarious examples can include AI-assisted computer-hacking and lethal autonomous weapons systems (LAWS), a.k.a. “killer robots.” Additional fears, of varying degrees of plausibility, include scenarios like those in the movies “2001: A Space Odyssey,” “Wargames,” and “Terminator.” While movies and weapons technologies might seem to be extreme examples of how AI might empower evil, we should remember that competition and war are always primary drivers of technological advance, and that militaries and corporations are working on these technologies right now. History also shows that great evils are not always completely intended (e.g., stumbling into World War I and various nuclear close-calls in the Cold War), and so having destructive power, even if not intending to use it, still risks catastrophe. Because of this, forbidding, banning, and relinquishing certain types of technology would be the most prudent solution.

### Economy

#### AI has broad societal and economic consequences – impact entire societies

Verdegem 22 , P. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. AI & Soc (2022). <https://doi.org/10.1007/s00146-022-01437-8> Dr. Pieter Verdegem is a Senior Lecturer in Media Theory in the Westminster School of Media and Communication and a member of [CAMRI](https://camri.ac.uk/) (the Communication and Media Research Institute). – Maren Lien

An important aspect of understanding AI capitalism is to consider AI as a *General Purpose Technology* (GPT) (Trajtenberg [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR69)). GPT are enabling technologies, meaning that they open up new opportunities, in addition to offering complete, final solutions. Other examples of GPT are the steam engine, electrification and the Internet. They have three main characteristics: (1) they are widely used; (2) they are capable of ongoing technical improvement; and (3) they enable innovation in different application sectors (Bresnahan, 210: 764). AI qualifies this definition. Given their pervasiveness and the complementary waves of innovation they produce, GPT cause economic disruption. They affect **entire economies**, potentially **drastically altering societies** through their **impact** on pre-existing economic and **social structures** (Trajtenberg [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR69)). Economists study the impact of GPT in terms of the emergence of winners and losers. The *winners* are those associated with the emerging GPT, whereas the *losers* are those who cannot benefit from the unfolding GPT. However, looking at other GPT invites us to look beyond winners and losers and to consider GPT—and thus also AI—as a public utility. The importance of electricity and the Internet, for example, has opened debates about the need of regulation, and the decision to not merely leaving these technological developments over to the market alone, or at least to have some intervention from society in it. This is particularly important given the times we live in: during the pandemic, we all have witnessed the crucial role of digital platforms in everyday life. As such, we need to be aware that AI can facilitate a further **polarisation** of already **unequal** societies (Crawford [2021](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR22); Dyer-Witheford et al. [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR27); Lee [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR41)). In any case, considering AI as a GPT and not just a digital technology that is owned and used by private entities but one that has broad impact on society, opens up new questions about how to conceptualise AI capitalism. This is where the work of Kate Crawford comes into place. In her book *Atlas of AI,* Crawford ([2021](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR22)) offers a comprehensive and nuanced understanding of AI. According to her, AI simultaneously refers to *technical approaches*, *social practices* and *industrial infrastructures* (Crawford [2021](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR22): 8–9). First, AI refers to technical approaches. Advancements in *machine learning* (ML) have been the most powerful contributor to the development of AI in the past two decades (Asaro [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR5)). ML is a paradigm that allows programs to automatically improve their performance on a particular task by learning from vast amounts of data (Russell and Norvig [2016](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR56); Lee [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR41)). It is based on statistical patterns and correlation in large data sets, starting to be used in the late 1980s–early 1990s. Earlier versions of machine intelligence—e.g., expert systems—were primarily rules-based, making use of symbolic logic and involving human experts generating instructions codified as algorithms (Agrawal et al. [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR1)). The problem was that they could not cope with the complexity of most applications. Unlike expert systems, powerful ML algorithms learn from the ground up, not from humans but from data (Alpaydin [2017](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR2)). The rise of ML can be explained by more powerful and reliable computing infrastructure, which has made possible the development of systems driven by real-world data (Lee [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR41)). The availability of significant amounts of data further enables the development of learning algorithms that derive solutions using statistical methods. *Deep learning* (DL) and *neural networks* (NN) are the driving forces behind more recent developments in ML. In the early 2000s, ML pioneer Geoffrey Hinton (LeCun et al. [2015](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR40)) demonstrated the power of DL neural networks: this allows automatically processing of unlabelled data, which has led to more effective applications of AI that we are now using every day (e.g., online services). Second, the *social practices* of AI refer to the classification systems, developed by humans, which are behind algorithms, ML/DL models and AI systems. Crucial questions that we need to ask here are: **Who** is involved in developing these **classification** systems? Who **decides** what classifications are used? and; What do they **look** like? Ultimately, these are **political** (power) questions about **inclusion** and **representation** (Crawford [2021](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR22)). Important **challenges** exist around AI, **bias**, **fairness** and **discrimination** (Costanza-Chock [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR18)). Questions about how representative these classification systems are, are crucial in this. How to avoid bias and support inclusion in AI systems are important political issues that urgently need to be addressed (Brevini and Pasquale [2020](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR12)). Last, the ***industrial infrastructures*** of **AI** refer to the computing power, algorithms and data sets that are the source of knowledge and production. This infrastructure not only entails the possibilities of collecting vast amounts of data—which are **need**ed to train algorithms—but also the computational power necessary to develop and perform ML and DL models. **Few** companies have **access** to the **required** data sets, possess the **necessary computational power** to run ML/DL and are able to attract the brightest AI scientists, which means we are witnessing a **concentrated** industrial **AI** infrastructure, leading to AI **oligopolies/monopolies** (Dyer-Witheford et al. [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR27); Riedl [2020](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR55)). This gives a lot of **power** in the hands of a **small** number of **corporations** (Montes and Goertzel [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR48)) and is why we need to scrutinise **economic power** within **AI capitalism**. Offering an encompassing view of AI capitalism, is important: we need to be aware of how material AI is and that its **production** is based on **natural resources, human labour** and **industrial infrastructures**. Looking at the broader picture of change within technologies, beliefs and infrastructures simultaneously, however, also risks overlooking the issues of a concentration of power. To deal with this, we need to go back to political economy as this is the framework that puts power at the centre of its analysis. Political economy is particularly interested in the relationship between techno-economic systems and their impact of the broader societal structure (McChesney [2000](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR45)). The industrial infrastructures of AI also contribute to a **concentration of powe**r, which has not only an **impact** on the **social practices of AI** but also how its **technological development** will happen in the **future**, which explains the importance of this perspective.

#### Economic positive feedback loop of AI gives powerful companies more power

Verdegem 22 , P. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. AI & Soc (2022). <https://doi.org/10.1007/s00146-022-01437-8> Dr. Pieter Verdegem is a Senior Lecturer in Media Theory in the Westminster School of Media and Communication and a member of [CAMRI](https://camri.ac.uk/) (the Communication and Media Research Institute). – Maren Lien

*Commodification* is a central concept in CPE and refers to the processes, whereby online and offline objects, activities, ideas and emotions are transformed into tradable commodities, transforming use value into exchange value (Hardy [2014](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR33)). In the context of AI capitalism, commodification is closely linked to *datafication*. The latter concept refers to the ability to render into data many aspects of the world that have never been quantified before (Cukier and Mayer-Schoenberger [2013](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR23)). Our social relationships, communication patterns, shopping behaviour, etc. are transformed into digital data (Couldry and Mejias [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR19)), which is an essential characteristic of the attention economy (Wu [2017](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR77)). In AI capitalism, the interplay between data and digital platforms is important. *Platforms* are intermediaries that invite different types of users—producers and suppliers, consumers, advertisers, app developers, etc.—to engage and interact via their digital infrastructure (Srnicek, [2017](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR60); Van Dijck et al. [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR71)). Platforms are ideally positioned to function as a data broker: central in their business model is the possibility to capture, extract and analyse the data produced by the interactions on the platform (Crain [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR21); West [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR76)). Using this extracted data as well as the skills workers gained when analysing it, made platform companies the leaders in the digital economy; working with data has become ever more important for gaining a competitive advantage (Srnicek [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR61)). What connects data and platforms are *network effects*. Network effects mean that the value of the network is determined by its size (Katz and Shapiro [1985](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR38)). Platforms thus become more valuable as more users join it. Engagement and interaction are only possible if there are active users on platforms. Generating network effects is thus a key strategic focus for platforms (Srnicek [2017](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR60)). The power of network effects goes hand in hand with the availability of data: this combination further strengthens the leading position of already powerful data companies (Srnicek [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR61)). Data-driven network effects entail that more users active on a certain platform, means more possibilities for data collection, analysis and extraction. Consequently, this results in more opportunities to use that data for improving the features and services offered by the platform. Better services open up the possibility to attract more users. A similar positive *data feedback loop* exists for AI too: better access to data means more opportunities to train ML models and better AI also results in better services and more users (Lee [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR41); Srnicek [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR61); Varian [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR72))

#### AI capitalism commodifies extracted data to reward dirty business practices of the powerful // OR (LINK) AI economies grows broader capitalistic economic system to reward and favor dirty business practices of the powerful

Verdegem 22, P. Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. AI & Soc (2022). <https://doi.org/10.1007/s00146-022-01437-8> Dr. Pieter Verdegem is a Senior Lecturer in Media Theory in the Westminster School of Media and Communication and a member of [CAMRI](https://camri.ac.uk/) (the Communication and Media Research Institute). – Maren Lien

While **data** is often considered as a raw material or a commodity, it makes sense to conceptualise it as a form of **capital** too. This is part of a broader discussion about how value is generated in the **contemporary economy** (Arvidsson and Colleoni [2012](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR4); Mazzucato [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR43)), particularly how **value** is derived from **data** and what normative aspects are relevant in the context of data collection and extraction (Couldry and Mejias [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR19); Mezzadra and Neilson [2017](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR47); Zuboff [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR78)). Sadowski ([2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR57)) argues that treating data as capital allows for a more nuanced and detailed understanding of how AI capitalism functions and is organised. What is the problem with using data to create value, as a resource to develop and optimise AI systems? Mazzucato ([2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR43)) analyses contemporary capitalism and highlights the critique that it **rewards** ***rent seekers*** over **true *value creators***. Their rent seeking is based on **overcharging prices**, **undercutting competition**—by **exploiting** particular advantages, e.g., **labour**, or using a **monopoly** advantage. Where *value creation* refers to the use of different types of resources to produce new goods and services, *value extraction* is defined as “*activities focused on moving around existing resources and outputs, and gaining disproportionally from the ensuing trade*” (Mazzucato [2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR43): 6). Data extraction is a particular type of value extraction. Sadowski ([2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR57): 9) defines *data extraction* as: “data is taken without meaningful consent and fair compensation for the producers and sources of data”*.* Evgeny Morozov ([2018](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR49)) follows a similar line of thinking and has coined *data extractivism* to refer to practices of tech giants launching products not for the revenue but for the data, which is afterwards monetised through different products and services (see also Couldry and Mejias [2019](https://link.springer.com/article/10.1007/s00146-022-01437-8#ref-CR19)). It is clear we must scrutinise what the **consequences** are of data **commodification** and **extraction** in **AI capitalism** as well as considering alternatives.

### Environment

#### AI fuels environmental degradation and leads to an ethics decline

Santa Clara University ND Santa Clara University, No date, "Artificial Intelligence and Ethics: Sixteen Challenges and Opportunities," No Publication, <https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics-sixteen-challenges-and-opportunities/> [AJL]

8. Environmental Effects Machine learning models require enormous amounts of energy to train, so much energy that the costs can run into the tens of millions of dollars or more. Needless to say, if this energy is coming from fossil fuels, this is a large negative impact on climate change, not to mention being harmful at other points in the hydrocarbon supply chain. Machine learning can also make electrical distribution and use much more efficient, as well as working on solving problems in biodiversity, environmental research, resource management, etc. AI is in some very basic ways a technology focused on efficiency, and energy efficiency is one way that its capabilities can be directed. On balance, it looks like AI could be a net positive for the environment [6]—but only if it is actually directed towards that positive end, and not just towards consuming energy for other uses. 9. Automating Ethics One strength of AI is that it can automate decision-making, thus lowering the burden on humans and speeding up – potentially greatly speeding up—some kinds of decision-making processes. However, this automation of decision making will presents huge problems for society, because if these automated decisions are good, society will benefit, but if they are bad, society will be harmed. As AI agents are given more powers to make decisions, they will need to have ethical standards of some sort encoded into them. There is simply no way around it: the ethical decision-making process might be as simple as following a program to fairly distribute a benefit, wherein the decision is made by humans and executed by algorithms, but it also might entail much more detailed ethical analysis, even if we humans would prefer that it did not—this is because Ai will operate so much faster than humans can, that under some circumstances humans will be left “out of the loop” of control due to human slowness. This already occurs with cyberattacks, and high-frequency trading (both of which are filled with ethical questions which are typically ignored) and it will only get worse as AI expands its role in society. Since AI can be so powerful, the ethical standards we give to it had better be good. 10. Moral Deskilling & Debility If we turn over our decision-making capacities to machines, we will become less experienced at making decisions. For example, this is a well-known phenomenon among airline pilots: the autopilot can do everything about flying an airplane, from take-off to landing, but pilots intentionally choose to manually control the aircraft at crucial times (e.g., take-off and landing) in order to maintain their piloting skills. Because one of the uses of AI will be to either assist or replace humans at making certain types of decisions (e.g. spelling, driving, stock-trading, etc.), we should be aware that humans may become worse at these skills. In its most extreme form, if AI starts to make ethical and political decisions for us, we will become worse at ethics and politics. We may reduce or stunt our moral development precisely at the time when our power has become greatest and our decisions the most important. This means that the study of ethics and ethics training are now more important than ever. We should determine ways in which AI can actually enhance our ethical learning and training. We should never allow ourselves to become deskilled and debilitated at ethics, or when our technology finally does present us with hard choices to make and problems we must solve—choices and problems that, perhaps, our ancestors would have been capable of solving—future humans might not be able to do it. For more on deskilling, see this article [7] and Shannon Vallor’s original article on the topic [8].

## Alts

### Decolonial AI

#### **Decolonial AI challenges US techno benevolence and opens space for sociotechnical interventions and affective community building**

Mohamed et al. 20 (Shakir Mohamed is Senior Staff Scientist at DeepMind, Marie Therese, William Isaac, 07-12-20, “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence” <https://link.springer.com/content/pdf/10.1007/s13347-020-00405-> pp. 669-671 pp. 671)-qcl

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.

### Corporeality

#### There is a distinction between the ‘body and ‘embodiment’, human and computer, but technology connects them – posthumanism loops them both

Wilcox ’16 (Lauren Wilcox, PhD, is a Staff Research Scientist in People & AI Research.--“Embodying algorithmic war: Gender, race, and the posthuman in drone warfare”—Article--Volume 48 Issue 1, February 2016--https://journals.sagepub.com/doi/abs/10.1177/0967010616657947)//Marzz

Hayles’s distinction between ‘body’ and ‘embodiment’ is particularly useful here: while ‘body’ is an abstract, idealized, universal construct, ‘embodiment’ is contextual, fully imbricated in culture, and never quite complies with the abstract idea of a ‘body’. This notion of incorporation is borne out in Hayles’s (1999) genealogy of cybernetics, which describes a shift from artificial intelligence to artificial life. The famous ‘Turing test’ is representative of the first wave, which attempted to divest information processing from a human body. This test, described by Alan Turing, suggests that the answer to whether or not machines can ‘think’ can be ascertained in an ‘imitation game’ in which a human poses questions at a computer terminal. If the human cannot tell whether the answers are generated by a human or a computer, then the computer can ‘think’. **The machine’s ability to ‘think’ is based on its ability to imitate a human**. While divesting information from its body, the Turing test also points toward cyborg subjectivity as it separates enacted bodies and represented bodies (through the terminal) while bringing them together in the technology that connects them. The second wave of cybernetics associated with artificial life sought to redefine and reorganize the boundaries of the human body in and through imbrication in technology and in relation to other bodies. Artificial life, in contrast to artificial intelligence, defines the human in terms of the machine: humans are understood as information processors that evolved intelligence. The contrast is between machines constructed to imitate humans (as in the Turing test) and what Hayles describes as ‘the computational model’ or ‘regime of computation’ in which all life, including humans, is understood as a kind of self-organizing machine. Human subjectivity is understood in relation to the machine. Both of these models exist in the figure of the posthuman, in which feedback loops between the human and the machine mutually inform and construct one another. Both the bodies of drone pilots as well as the bodies of those targeted by such algorithmic regimes exceed or overflow their constitution in this ‘regime of computation’, and as such point to the ways in which bodies as **posthuman bodies must be considered in terms of multiple sources of corporealization**. Thus, the use of algorithms and ‘big data’ techniques to produce targetable bodies is a process that cannot be separated from the greater process in which the drone assemblage is made of posthuman bodies: bodies corporealized as processors of information and also as information bundles. While this form of corporealization is a key feature of drone warfare, the massacre in Uruzgan province reveals that the visual and affective registers of embodiment are also necessary and are particularly crucial in the embodiment of gender and race in drone warfare.

#### Posthuman feminist work analyze different bodies through means of differentiation—questions the process of becoming

Wilcox ’16 (Lauren Wilcox, PhD, is a Staff Research Scientist in People & AI Research.--“Embodying algorithmic war: Gender, race, and the posthuman in drone warfare”—Article--Volume 48 Issue 1, February 2016--https://journals.sagepub.com/doi/abs/10.1177/0967010616657947)//Marzz

The work of posthuman feminists provides a necessary check on tendencies to theorize the drone as ‘other than human’ in ways that reinforce the separation of humans from techno-scientific practices, including the use of visual technologies, algorithms, and artificial intelligence in various configurations to enable ‘drone warfare’. The challenge the posthuman body poses is not the addition of new technological advances to an already-existing human body, but rather the addition of a body that is always already formed through norms and relations to others, whether these other are human, technological, or animal (see also Braidotti, 2013). Notably, N Katherine Hayles critiques the concept of the posthuman for its rehabilitation of a liberal subject of autonomy and individuality in the figure of the human. For Hayles, the posthuman ‘signals…the end of a certain conception of the human, a conception that may have applied, at best, to that fraction of humanity who had the wealth, power and leisure to conceptualize themselves as autonomous beings exercising their will though individual agency and choice’ (Hayles, 1999: 286). As such, posthuman feminist projects critique a kind of ‘transhumanism’ that seeks to disembody consciousness or promote ‘other than’ or ‘more than’ human approaches that reify a particular normative version of humanity that enables distinctions between more or less worthy forms of life. Relatedly, posthuman feminist approaches take seriously the ways in which embodied differences are produced and lived, avoiding totalizing visions that might obscure the ways in which bodies are differentially produced through the interaction of technological processes with gendered, racial, colonialist, sexual, and other means of differentiation that are themselves emergent processes of identification and alignment with other bodies. As such, I understand the turn toward data and machine intelligence not as an ‘other than-human’ process of decision-making but as a form of embodiment that reworks and undermines essentialist notions of culture and nature, biology and technology, often but not necessarily in the service of projects of domination. Haraway’s concept of corporealization as ‘the interactions of humans and nonhumans in the distributed, heterogeneous work processes of technoscience’ (1997: 141) is highly relevant to theorizing drone warfare. Post-9/11 practices of security have intensified what Haraway describes as ‘corporeal fetishism’ (Haraway, 1997: 142) in which the sciences are imagined to be able to simplify and condense complex relationships and situated knowledges into singular digital maps of the body that are free from the alleged ‘failures’ of culture, but are in fact themselves corporealizing practices (see Pötzsch, 2015 for other examples). Following Haraway, I ask not to what extent drone warfare is embodied or disembodied, but rather, how bodies are corporealized in drone assemblages with a caveat that the corporealization is always in a process of ‘becoming’. The first aspect of corporealization I analyze is the production of targetable, killable bodies in drone assemblages.

#### **Encoding gender reproduces logics of disposability and authorizes war power against “hated” bodies – blurring the lines between artificial intelligence and embodiment is key to disrupting the naturalization of gender and racial formations**

Wilcox ’16 (Lauren Wilcox, PhD, is a Staff Research Scientist in People & AI Research.--“Embodying algorithmic war: Gender, race, and the posthuman in drone warfare”—Article--Volume 48 Issue 1, February 2016--https://journals.sagepub.com/doi/abs/10.1177/0967010616657947)//Marzz

In an early essay with contemporary resonances to the Black Lives Matter movement in the USA, Judith Butler analyzes the video-taped beating of Rodney King in Los Angeles in the 1990s and the acquittal (despite the video evidence) of the police who beat King as indicating that ‘the visual field is not neutral to the question of race; it is itself a racial formation, and episteme, hegemonic and forceful’ and therefore ‘seeing’ is thus not a matter of direct perception as a means to truth but ‘the racial production of the visible, the workings of racial constraints on what it means to “see”’ (Butler, 1993: 16). An ‘inverted projection of white paranoia’ posited the object of violence as subject of violence. This case also speaks to the blurred lines between the war power and police powers of drone warfare and the interventions in Iraq and Afghanistan and the use of violence against racialized bodies (Neocleous, 2014; Holmqvist, 2014). The designation of prayer as a signal of intent to do something ‘nefarious’, the movements of a vehicle as a ‘flanking maneuver’, the interpretation of movement of bodies in the back of a pick-up trick as the use of ‘human shields’ (not that these ‘human shields’ prevented the bombing of the vehicle), and perceived presence of only men and only adolescents or older suggests that greater accuracy of vision is unlikely to serve as a check on the mistakes of either algorithmic or visual analysis when the bodies are already perceived, or rather, felt to be dangerous. The logic of killing based not on identity but on these visual cues taken to resemble the algorithmic ‘pattern of life’ bears a striking resemblance to performative theories of gender in which ‘gender’ is not an essential identity one embodies; rather, one comes to be embodied as a subject through **repeated behaviors** and practices **against a normative framework that renders some modes of life into the category of ‘hated’**. J Halberstam writes, ‘**Gender**, we might argue, **like computer intelligence, is a learned**, imitative behavior that can be processed **so well that it comes to look natural’** (Halberstam, 1991: 443). Halberstam purposefully blurs the line between artificial intelligence and embodied behavior in conjunction with Turing’s ‘other’ test. Somewhat less famous is Turing’s corollary test in which a person interacts over a terminal with a man and a woman in a different room, trying to ascertain from people who may be attempting to deceive the tester based on written language as to which respondent is embodied as a man or a women. While the first test simultaneously posits the possibility of ‘thinking’ without a (human) body (while constructing a cyborg whose represented and enacted bodies are linked via technology), the ‘gender’ Turing test opens up the possibility of failure to unite represented gender embodiment and enacted gender embodiment into a single identity. As Hayles writes, What the Turing test ‘proves’ is that the overlay between the enacted and the represented bodies is no longer a natural inevitability but a contingent production, mediated by a technology that has become so entwined with the production of identity that it can no longer meaningfully be separated from the human subject. (Hayles, 1999: xiii) Gender proves to be a feature of human embodiment that must be represented and read via different technologies, the reading of which will prove crucial for the process of racialization that makes certain bodies killable.

## AT Perm

#### Perm fails - US is teaming up with Russia to block AI weapon regulations

David H. Freedman, 9/15/21, “US Is Only Nation with Ethical Standards for AI Weapons. Should We Be Afraid?”, Newsweek, David H. Freedman is a scientific journalist, author, and is a contributing writer at The Atlantic and Newsweek, <https://www.newsweek.com/2021/09/24/us-only-nation-ethical-standards-ai-weapons-should-we-afraid-1628986.html> - Maren Lien

That sets up a **race-to-the-bottom** in which the **least ethical** or most **careless** adversary—one that is most aggressive about fielding AI-enabled weaponry, regardless of reliability and safeguards—**forces** others to follow suit. Nuclear weapons could be placed under the control of flawed AI systems that watch for signs that someone else's AI nukes are about to launch. AI is "increasing the risk of inadvertent or accidental escalation caused by misperception or miscalculation," says James Johnson, a foreign-policy researcher at Ireland's Dublin City University and author of Artificial Intelligence and the Future of Warfare. (Manchester University Press, September 2021). Both the **U.S**. and **Russia** have repeatedly **refused** to allow the **U**nited **N**ations' Convention on Certain Conventional Weapons (CCW), the main international body for weapons agreements, to **ban lethal AI-controlled weapons**. Meetings to discuss revisiting the CCW are planned for December, but there's **little** optimism an agreement will be reached; among the most powerful nations, only China has expressed support for such a treaty. [NATO](https://www.newsweek.com/topic/nato) nations have discussed the possibility of an agreement, but nothing definite has emerged. If the **U.S**. is negotiating **AI** weapons **separately** with other countries, there's **little** public word of it. Even if diplomatic efforts led to limits on the use of AI, verifying adherence would be far more difficult than, say, inspecting nuclear missile silos. **Military leaders** in a **hostile**, competitive world are not known for their ability to **resist** advanced **weaponry**, regardless of **consequences**.

# AFF ANSWERS

## AT Link

### Governance/Regulation

#### Regulation is key to capture the benefits of AI – solves laundry list of societal and security issues

Tzimas ’21 [Themistoklis; 2021; Faculty of Law at the Aristotle University of Thessaloniki; Legal and Ethical Challenges of Artificial Intelligence from an International Law Perspective, “Chapter 2: The Expectations and Risks from AI,” p. 9-32]

Therefore, it is only natural to be at least skeptical towards a future with entities possessing equal or superior intelligence and levels of autonomy; the prospect even of existential risk looms as possible.7 AI that will have reached or surpassed our level of intelligence make us wonder why would highly autonomous and intelligent AI want to give up control back to its original creators?8 Why remain contained in pre-deﬁned goals set for it by us, humans?Even AI in its current form and narrow intelligence poses risks because of its embedded-ness in an ever-growing number of crucial aspects of our lives. The role of AI in military, ﬁnancial,9 health, educational, environmental, governance networks-among others—are areas where risk generated by AI—even limited— autonomy can be diffused through non-linear networks, with signiﬁcant impact— even systemic.10 The answer therefore to the question whether AI brings risk with it is yes; as Eliezer Yudkowski comments the greatest of them all is that people conclude too early that they understand it11 or that they assume that they can achieve it without necessarily having acquired complete and thorough understanding of what intelli- gence means.12 Our projection of our—lack of complete—understanding of the concept of intelligence on AI is owed to our lack of complete comprehension of human intelligence too, which is partially covered by the prevalent and until now self- obvious, anthropomorphism because of which we tend to identify higher intelligence with the human mind.Yudkowski again however suggests that AI “refers to a vastly greater space of possibilities than does the term “Homo sapiens.” When we talk about “AIs” we are really talking about minds-in-general, or optimization processes in general. Imagine a map of mind design space. In one corner, a tiny little circle contains all humans; within a larger tiny circle containing all biological life; and all the rest of the huge map is the space of minds-in-general. The entire map ﬂoats in a still vaster space, the space of optimization processes.”13 Regardless of what our well-established ideas are, there are many, different intelligences and even more signiﬁcantly, there are potentially, different intelli- gences equally or even more evolved than human.From such a perspective, the unprecedented—ness of potential AI developments and the mystery surrounding them emerges as not only the outcome of pop culture but of a radical transformation of our—until recently—self—obvious identiﬁcation of humanity with highly evolved and dominant intelligence.14 The lack of understanding of intelligence and therefore of AI may be frightening but does not lead necessarily to regulation—at least to a proper one. We could even be led into making potentially catastrophic choices, on the basis of false assumptions. On top of our lack of understanding, we should add a sentiment of anxiety as well as of expectations, which intensiﬁes as an atmosphere of emergency and of expected groundbreaking developments grows. The most graphic description of this feeling is the potential of a moment of singularity, as mentioned above according to the description by Vinge and Kurzweil. As the mathematician I. J. Good–Alan Turing’s colleague in the team of the latter during World War II—has put it: “Let an ultraintelligent machine be deﬁned as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the ﬁrst ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.”15 This is in a nutshell the moment of singularity. The estimates currently foresee the emergence of ultra or super intelligence—as it is currently labelled—or in other words of singularity, somewhere between 20 and 50 years from today, further raising the sentiment of emergency.16 We cannot even foretell with precision how singularity would look like but we know that because of its expected groundbreaking impact, both states and private entities compete towards gaining the upper hand in the prospect of the singularity.17 Despite the fact that such predictions have been proven rather optimistic in the past18 and therefore up to some extent inaccurate, there are reasons to assume that their materialization will take place and that the urgency of regulation will be proven realistic. After all, part of the disappointments from AI should be blamed on the fact that certain activities and standards, which were considered as epitomes of human intelligence have been surpassed by AI, only to indicate that they were not eventu- ally satisfactory thresholds for the surpassing of human intelligence.19 Partially because of AI progress we realize that human intelligence and its thresholds are much more complicated than assumed in the past. The vastness’s of deﬁnitions of intelligence, as well as its etymological roots are enlightening of the difﬁculties: “to gather, to collect, to assemble or to choose, and to form an impression, thus leading one to ﬁnally understand, perceive, or know”.20 As with other relevant concepts, the truth is that until recently our main way to approach intelligence for far too long was “we know it, when we see it”. AI is an additional reason for looking deeper into intelligence and the more we examine it, the most complicated it seems. The combination of lack of complete understanding of intelligence, the unpredictability of AI, its rapid evolution and the prospect of singularity explain both the fascination and the fear from AI. Once the latter emerges, we have no real knowledge about what will happen next but only speculations, which until recently belonged to the area of science ﬁction. We are for example pretty conﬁdent that the speed of AI intelligence growth will accelerate, once self—improvement will have been achieved. The expected or possible chain of events will begin from AI capacity to re-write its own algorithms and exponentially self—improve, surpassing human intelligence, which lacks the capacity of such rapid self—improvement and setting its own goals.21 We can somehow guess the speed of AGI and ASI evolution and possibly some of its initial steps but we cannot guess the directions that such AI will choose to follow and the characteristics that it will demonstrate. Practically, we credibly guess the prospects of AI beyond a certain level of development. Two existential issues could emerge: ﬁrst, an imbalance of intelligence at our expense—with us, humans becoming the inferior species—in favor of non-biological entities and secondly a lack of even fundamental conceptual communication between the two most intelligent “species”. Both of them heighten the fear of irreversible changes, once we lose the possession of the superior intelligence.22 However, we need to consider the expectations as well. The positive side focuses on the so-called friendly AI, meaning AI which will beneﬁt and not harm humans, thanks to its advanced intelligence.23 AI bears the promise of signiﬁcantly enhancing human life on various aspects, beginning from the already existing, narrow applications. The enhanced automation24 in the industry and the shift to autonomy,25 the take—over by AI of tasks even at the service sector which can be considered as “tedious”—i.e. in the banking sector—climate and weather forecasting, disaster response,26 the potentially better cooperation among different actors in complicated matters such as in matters of information, geopolitics and international relations, logistics, resources ex.27 The realization of the positive expectations depends up to some extent upon the complementarity or not, of AI with human intelligence. However, what friendly AI will bring in our societies constitutes a matter of debate, given our lack of unanimous approach on what should be considered as beneﬁcial and therefore friendly to humans—as is analyzed in the next chapter.Friendly AI for example bears the prospect of freeing us from hard labor or even further from unwanted labor; of generating further economic growth; of dealing in unbiased, speedy, effective and cheaper ways with sectors such as policing, justice, health, environmental crisis, natural disasters, education, governance, defense and several more of them which necessitate decision-making, with the involvement of sophisticated intelligence.The synergies between human intelligence and AI “promise” the enhancement of humans in most of their aspects. Such synergies may remain external—humans using AI as external to themselves, in terms of analysis, forecasts, decision—making and in general as a type of assistant-28 or may evolve into the merging of the two forms of intelligence either temporarily or permanently.The second profoundly enters humanity, existentially—speaking, into uncharted waters. Elon Musk argues in favor of “having some sort of merger of biological intelligence and machine intelligence” and his company “Neuralink” aims at implanting chips in human brain. Musk argues that through this way humans will keep artiﬁcial intelligence under control.29 The proposition is that of “mind design”, with humans playing the role that God had according to theologies.30 While the temptation is strong—exceeding human mind’s capacities, far beyond what nature “created”, by acquiring the capacity for example to connect directly to the cyberspace or to break the barriers of biology31—the risks are signiﬁcant too: what if a microchip malfunction? Will such a brain be usurped or become captive to malfunctioning AI?The merging of the two intelligences is most likely to evolve initially by invoking medical reasons, instead of human enhancement. But the merging of the two will most likely continue, as after all the limits between healing and enhancement are most often blurry. This development will give rise, as is analyzed below, to signif- icant questions and issues, the most of crucial of which is the setting of a threshold for the prevalence of the human aspect of intelligence over the artiﬁcial one.Human nature is historically improved, enhanced, healed and now, potentially even re-designed in the future.32 Can a “medical science” endorsing such a goal be ethically acceptable and if yes, under what conditions, when, for whom and by what means? The answers are more difﬁcult than it seems. As the World Health Organi- zation—WHO—provides in its constitution, “Health is a state of complete physical, mental and social well-being and not merely the absence of disease or inﬁrmity”.33 Therefore, why discourage science which aims at human-enhancement, even reaching the levels of post-humanism?34 Or if restrictions are to be imposed on human enhancement, on what ethics and laws will they be justiﬁed? How ethically acceptable is it to prohibit or delay technological evolution, which among several other magniﬁcent achievements, promises to treat death as a disease and cure it, by reducing soul to self, self to mind, and mind to brain, which will then be preserved as a “softwarized” program in a hardware other than the human body?35 After all, “According to the strong artiﬁcial intelligence program there is no fundamental difference between computers and brains: a computer is different machinery than a person in terms of speed and memory capacity.”36 While such a scientiﬁc development and the ones leading potentially to it will be undoubtedly, groundbreaking technologically-speaking, is it actually—ethically- speaking—as ambivalent as it may sound or is it already justiﬁed by our well— rooted human-centrism?37 Secular humanism may have very well outdated religious beliefs about afterlife in the area of science but has not diminished the hope for immortality; on the contrary, science, implicitly or explicitly predicts that matter can in various ways surpass death, albeit by means which belong in the realm of scientiﬁc proof, instead of that of metaphysical belief.38 If this is the philosophical case, the quest for immortality becomes ethically acceptable; it can be considered as embedded both in the existential anxiety of humans, as well as in the human-centrism of secular philosophical and political victory over the dei-centric approach to the world and to our existence. From another perspective of course and for the not that distant philosophical reasons, the quest for immortality becomes ethically ambiguous or even unacceptable.39 By seeking endless life we may miss all these that make life worth living in the framework of ﬁniteness. As the gerontologist Paul Hayﬂick cautioned “Given the possibility that you could replace all your parts, including your brain, then you lose your self-identity, your self-recognition. You lose who you are! You are who you are because of your memory.”40 In other words, once we begin to integrate the two types of intelligence, within ourselves, until when and how we will be sure that it is human intelligence that guides us, instead of the AI? And if we are not guided completely or—even further—at all by human intelligence but on the contrary we are guided by AI which we have embodied and which is trained by our human intelligence, will we be remaining humans or we will have evolved to some type of meta-human or transhumant species, being different persons as well?41 AI promises tor threatens to offer a solution by breaking down our consciousness into small “particles” of information—simplistically speaking—which can then be “software-ized” and therefore “uploaded” into different forms of physical or non-physical existence. Diane Ackerman states that “The brain is silent, the brain is dark, the brain tastes nothing, the brain hears nothing. All it receives are electrical impulses--not the sumptuous chocolate melting sweetly, not the oboe solo like the ﬂight of a bird, not the pastel pink and lavender sunset over the coral reef--only impulses.”42 Therefore, all that is needed—although it is of course much more complicated than we can imagine—is a way to code and reproduce such impulses. Even if we consider that without death, we will no more be humans but something else, why should we remain humans once technologies allow us be something “more”, in the sense of an enhanced version of “being”? Why are we to remain bound by biological evolution if we can re-design it and our future form of existence? Why not try to achieve the major breakthrough, the anticipated or hoped digita- lization of the human mind, which promises immortality of consciousness via the cyberspace or artiﬁcial bodies: the uploading of our consciousness so that it can live on forever, turning death into an optional condition.43 Either through an artiﬁcial body or emulation-a living, conscious avatar—we hope—or fear—that the domain of immortality will be within reach. It is the prospect of a “substrate-independent minds,” in which human and machine consciousness will merge, transcending biological limits of time, space and mem- ory” that fascinates us.44 As Anders Sandberg explained “The point of brain emulation is to recreate the function of the original brain: if ‘run’ it will be able to think and act as the original,” he says. Progress has been slow but steady. “We are now able to take small brain tissue samples and map them in 3D. These are at exquisite resolution, but the blocks are just a few microns across. We can run simulations of the size of a mouse brain on supercomputers—but we do not have the total connectivity yet. As methods improve, I expect to see automatic conversion of scanned tissue into models that can be run. The different parts exist, but so far there is no pipeline from brains to emulations.”45 The emulation is different from a simulation in the sense that the former mimics not only the outward outcome but also the “internal causal dynamics”, so that the emulated system and in this particular case the human mind behaves as the original.46 Obviously, this is a challenging task: we need to understand the human brain with the help of computational neuroscience and combine simpliﬁed parts such as simulated neurons with network structures so that the patterns of the brain are comprehended. We must combine effectively “biological realism (attempting to be faithful to biology), completeness (using all available empirical data about the system), tractability (the possibility of quantitative or qualitative simulation) and understanding (producing a compressed representation of the salient aspects of the system in the mind of the experimenter)”.47 The technological challenges are vast. Technologically speaking, the whole concept is based on some assumptions which must be proven both accurate and feasible.48 We must achieve technology capable of scanning completely the human brain, of creating software on the basis of the acquired information from its scanning and of the interpretation of information and the hardware which will be capable of uploading or downloading such software.49 The steps within these procedures are equally challenging. Their detailed analysis evades the scope of this book. Some critical questions—they are further analyzed in the next chapters—emerge however: how will we interpret free will in emulation? What will be the impact of the environment and of what environment? How will be missing parts of the human brain re-constructed and emulated? What will be the status of the several emulations which will be created—i.e. failed attempts or emulations of parts of the human brain—in the course of the search for a complete and functioning emulation? Will they be considered as “persons” and therefore as having some right or will they be considered as mere objects in an experimental lab? How are we going to decode the actual subjective sentiments of these emulations? Essentially, are emulations the humans “themselves” who are emulated or a different person? Even further what will human and person mean in the era of emulation? From a different perspective, the victory over death may be seen as a danger of mass extinction, absorption or de-humanization. In this new, vast universe of emulations will there be place for humans?50 From the above—mentioned discussion, it becomes obvious that at a large extent, the prospect of risk or of expectation is a matter of perspective, for which there is no unanimous agreement in the present. This may be the greatest danger of all, for which Asimov warned us: unleashing technology while we cannot communicate among us, in the face of it. The existential prospect as well as the risks by AI may self-evidently emerge from technological advances but are determined on the basis of politico—philosophical or in the wider sense, ethical assumptions. This is where the need for legal regulation steps in. Such a need was often underestimated in the past in favor of a solely technologically oriented approach—although exceptions raising issues other than technological can be found too.51 The gradual raising of ethic—political, philosoph- ical and legal issues constitutes a rather recent development, partially because of the realization of the proximity of the risks and of the expectations. The public debate is often divided between two “contradictory” views: fear of AI or enthusiastic optimism. The opinions of the experts differ respectively. Kurzweil, who has come with a prediction for a date for the emergence of singularity—until 2045—expects such a development in a positive way: “What’s actually happening is [machines] are powering all of us,” Kurzweil said during the SXSW interview. “They’re making us smarter. They may not yet be inside our bodies, but, by the 2030s, we will connect our neocortex, the part of our brain where we do our thinking, to the cloud.”52 In a well-known article—issued on the occasion of a ﬁlm—Stephen Hawking, Max Tegmark, Stuart Russell, and Frank Wilczek shared a moderate position: “The potential beneﬁts are huge; everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magniﬁed by the tools AI may provide, but the eradication of war, disease, and poverty would be high on anyone’s list. Success in creating AI would be the biggest event in human history. . . Unfortunately, it might also be the last, unless we learn how to avoid the risks.”53

### AWS = Moral

#### Autonomous weapons better for ethics- reduces unnecessary killing

Galliott, Jai. "Humans, autonomous systems, and killing in war." Research Anthology on Military and Defense Applications, Utilization, Education, and Ethics. IGI Global, 2021. 240-257. [AJL]

However, not all are convinced by the argument that autonomous systems present a moral problem in making it easier to indiscriminately and disproportionately kill, even at the higher end of the spectrum of autonomy. Daniel Brunstetter and Megan Braun (2011, p. 339) argue that semi-autonomous robotic systems are subject to the same jus in bello requirements as other weapons used in war, but that their +XPDQV$XWRQRPRXV6\VWHPVDQG.LOOLQJLQ:DU technological advantages coupled with the removal of risk to soldiers means that they should, at the least in theory, make satisfying the principles of discrimination and proportionality an easier task and perhaps make operators more reluctant to kill in situations where doubt exists as to the legitimacy of the potential victim of aggression. They say that the in the case of surveillance, at the very least, the distance or what they call ‘separation factor’, arguably offers an increased level of control over lethal targeting decisions and ought to actually reduce the emotional toll and unnecessary killing (Brunstetter and Braun 2011, p. 339). They regard a drone operator’s ability to confer with a superior officer as being a critical factor encouraging ethical decision making in war. In some instances, this may be the case, and may even apply in the case of highly autonomous systems if the relevant coding and engineering is sufficiently detailed and comprehensive to account for all relevant morally relevant inputs and outputs, but in others, it might be that having a mission commander or a test overseeing the operator’s/programmer’s actions only places additional pressure on them to perpetrate lethal acts, just as the sergeants walking the trenches of WWI aimed to encourage reluctant soldiers to kill. Yet Christian Enemark (2013) also questions some of the assumptions relied on here. He says that there is reason to suppose that being physically absent from the battlefield is more conducive to discrimination (Enemark 2013). In his view, the removal of risk allows decisions to be made in a more deliberate manner and also removes anger and emotion that he thinks might otherwise lead to morally unsanctioned killings. That is, if a drone operator working from a desk in Nevada encounters the enemy, adherence to jus in bello protocol should improve as the operator is at little or no personal risk. It could be suggested, however, that if an operator or technical contributors is so emotionally removed, they are in fact likely to develop the sort disengagement referred to above or an even more morally concerning callousness. In case of highly autonomous system with little input other than through code, the concern that it such a callousness might pervade said code and perhaps even go unnoticed by virtue of being concealed within the system and being evident only from its highly complex actions, having even more severe and long-lasting consequences.

#### Autonomous weapons are at best more ethical than humans and at worse just as ethical as humans

Gunkel 17 (David, Professor of Communication Studies at Northern Illinois University. 2017 “Mind the gap: responsible robotics and the problem of responsibility” Ethics and Information Technology. doi:10.1007/s10676-017-9428-2)-qcl

Conversely, we can entertain the possibility of what has been called “machine ethics” just as we had previously done for other non-human entities, like animals (Singer 1975). And there has, in fact, been a number of recent proposals addressing this opportunity. Wallach and Allen (2009, p. 4), for example, not only predict that “there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight” but use this fact as justification for developing “moral machines,” advanced technological systems that are able to respond to morally challenging situations. Anderson and Anderson (2011) take things one step further. They not only identify a pressing need to consider the moral responsibilities and capabilities of increasingly autonomous systems but have even suggested that “computers might be better at following an ethical theory than most humans,” because humans “tend to be inconsistent in their reasoning” and “have difficulty juggling the complexities of ethical decision-making” owing to the sheer volume of data that need to be taken into account and processed (Anderson and Anderson 2007, p. 5). These proposals, it is important to point out, do not necessarily require that we first resolve the “big questions” of AGI (Artificial General Intelligence), robot sentience, or machine consciousness. As Wallach (2015, p. 242) points out, these kinds of machines need only be “functionally moral.” That is, they can be designed to be “capable of making ethical determinations…even if they have little or no actual understanding of the tasks they perform.” The precedent for this way of thinking can be found in corporate law and business ethics. Corporations are, according to both national and international law, legal persons (French 1979). They are considered “persons” (which is, we should recall, a moral classification and not an ontological category) not because they are conscious entities like we assume ourselves to be, but because social circumstances make it necessary to assign personhood to these artificial entities for the purposes of social organization and jurisprudence. Consequently, if entirely artificial and human fabricated entities, like Google or IBM, are legal persons with associated social responsibilities, it would be possible, it seems, to extend the same moral and legal considerations to an AI or robot like Google’s DeepMind or IBM’s Watson. The question, it is important to point out, is not whether these mechanisms are or could be “natural persons” with what is assumed to be “genuine” moral status; the question is whether it would make sense and be expedient, from both a legal and moral perspective, to treat these mechanisms as persons in the same way that we currently do for corporations, organizations and other human artifacts. Once again, this decision sounds reasonable and justified. It extends both moral and legal responsibility to these other socially aware and interactive entities and recognizes, following the predictions of Wiener (1988, p. 16), that the social situation of the future will involve not just human-tohuman interactions but relationships between humans and machines and machines and machines. But this shift in perspective also has significant costs. First, it requires that we rethink everything we thought we knew about ourselves, technology, and ethics. It entails that we learn to think beyond human exceptionalism, technological instrumentalism, and many of the other -isms that have helped us make sense of our world and our place in it. In effect, it calls for a thorough reconceptualization of who or what should be considered a legitimate center of moral concern and why Second, robots that are designed to follow rules and operate within the boundaries of some kind of programmed restraint, might turn out to be something other than what is typically recognized as a responsible agent. Winograd (1990, pp. 182–183), for example, warns against something he calls “the bureaucracy of mind,” “where rules can be followed without interpretive judgments.” “When a person,” Winograd (1990, p. 183) argues, “views his or her job as the correct application of a set of rules (whether human-invoked or computerbased), there is a loss of personal responsibility or commitment. The ‘I just follow the rules’ of the bureaucratic clerk has its direct analog in ‘That’s what the knowledge base says.’ The individual is not committed to appropriate results, but to faithful application of procedures.” Coeckelbergh (2010, p. 236) paints a potentially more disturbing picture. For him, the problem is not the advent of “artificial bureaucrats” but “psychopathic robots.” The term “psychopathy” has traditionally been used to name a kind of personality disorder characterized by an abnormal lack of empathy which is masked by an ability to appear normal in most social situations. The functional morality, like that specified by Anderson and Anderson and Wallach and Allen, intentionally designs and produces what are arguably “artificial psychopaths”—robots that have no capacity for empathy but which follow rules and in doing so can appear to behave in morally appropriate ways. These psychopathic machines would, Coeckelbergh (2010, p. 236) argues, “follow rules but act without fear, compassion, care, and love. This lack of emotion would render them non-moral agents—i.e. agents that follow rules without being moved by moral concerns—and they would even lack the capacity to discern what is of value. They would be morally blind.”4 Efforts in “machine ethics” (or whatever other nomenclature comes to be utilized to name this development) effectively seek to widen the circle of moral subjects to include what had been previously excluded and marginalized as mere neutral instruments of human action. This is, it is important to note, not some blanket statement that would turn everything that was a tool into a moral subject. It is the recognition, following Marx, that not everything technological is reducible to a tool and that some devices—what Marx called “machines” and what Winner calls “autonomous technology”—might need to be programmed in such a way as to behave reasonably and responsibly for the sake of respecting human individuals and communities. This proposal has the obvious advantage of responding to moral intuitions: if it is the machine that is making the decision and taking action in the world with little or no direct human oversight, it would only make sense to hold it accountable (or at least partially accountable) for the actions it deploys and to design it with some form of constraint in order to control for possible bad outcomes. But doing so has considerable costs. Even if we bracket the questions of AGI, super intelligence, and machine consciousness; designing robotic systems that follow prescribed rules might provide the right kind of external behaviors but the motivations for doing so might be lacking. “Even if,” Sharkey (2012, p. 121) writes in a consideration of autonomous weapons, “a robot was fully equipped with all the rules from the Laws of War, and had, by some mysterious means, a way of making the same discriminations as humans make, it could not be ethical in the same way as is an ethical human. Ask any judge what they think about blindly following rules and laws.” Consequently, what we actually get from these efforts might be something very different from (and maybe even worse than) what we had hoped to achieve.

#### Autonomous weapons are moral – the ability for targets to fight back against autonomous weapons mean they respect the target’s autonomy

**Young 21** (Garry director at the GW Institute of Public Policy, 03-29-2021, accessed on 6-21-2022, Ethics and Information Technology, “On the indignity of killer robots. Ethics and Information Technology,” 23(3), 473–482. https://doi.org/10.1007/s10676-021-09590-2 , pp. 6

A strong rebuttal of the indignity argument denies the truth of P3 (that the deployment of killer robots disrespects the dignity of combatants). Positioning ourselves once more behind a veil of ignorance, we again ask: what would military commanders be agreeing to if they were to agree to the permissibility of killer robots? We know that they would be agreeing to deploy autonomous weapons whose decision making cannot be constrained by recognition respect. In deciding whether combatants live or die, the killer robots would be ‘treating’ them as objects and not as moral agents with inherent sortal dignity. In other words, they would be processing combatants in a manner that would be no diferent to any other object in their environment. Given this, the charge is that we (qua military commanders), by agreeing to deploy killer robots in this way, would be treating combatants in a manner that disrespect their sortal dignity. It is this claim that the strong rebuttal challenges. Permitting killer robots does not deny combatants the opportunity to fght back against these automated weapons, and therefore act as moral agents. The weapons themselves may not be capable of respecting the inherent dignity of the combatant they target, making their deaths appear arbitrary (Amoroso, 2017) but, in deciding to deploy such weapons, we are capable of recognizing the inherent dignity of the combatants these weapons will eventually target. After all, even if we accept that sortal dignity is inherent, what counts as an afront to this dignity is not immuable. Instead, it is constructed, and forms part of what Killmister (2017) refers to as social indignity. Therefore, from behind the veil of ignorance, we (qua the community of military commanders) could agree (socially construct and endorse the view) that killer robots are not an afront to sortal dignity because their deployment does not prevent combatants from acting as moral agents (exercising their rational autonomy) and quite possible neutralizing the killer robots in return: a fact we recognize and respect.11 Consequently, P3 of the indignity argument is false, meaning that C(ii) does not necessarily follow. Given the stronger rebuttal of the indignity argument, if P3 is rejected then C(iii)—the claim that the death of a combatant, as a consequence of P3, amounts to an undignifed death—is likewise rejected. In the case of the weaker rebuttal, however, where P3 is not rejected, might a case still be made for the truth of C(iii): that even when treating a lack of respect for the dignity of combatants as a pro tanto wrong, the afront to the combatant’s dignity nevertheless results in an undignifed death? I do not believe so, as I intend to show in the next section. By drawing on two examples from fction I defend the claim that one can preserve one’s outward dignity in the face of indignity but, also, that the preservation of dignity supports the stronger rebuttal’s claim that recognition respect would be bestowed on combatants faced with an assault from killer robots from the community of military commanders (as well as others), adding weight to the claim that the deployment of killer robots is not in fact undignifed. Either way, C(iii) is undermined.

### AI is fixable/ethical

#### AI could become more ethical than humans and provide more time for humans to benefit society

Bossmann 16 Julia Bossmann, 10-21-2016, "Top 9 ethical issues in artificial intelligence," World Economic Forum, <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> [AJL]

Optimizing logistics, detecting fraud, composing art, conducting research, providing translations: intelligent machine systems are transforming our lives for the better. As these systems become more capable, our world becomes more efficient and consequently richer. Tech giants such as Alphabet, Amazon, Facebook, IBM and Microsoft – as well as individuals like Stephen Hawking and Elon Musk – believe that now is the right time to talk about the nearly boundless landscape of artificial intelligence. In many ways, this is just as much a new frontier for ethics and risk assessment as it is for emerging technology. So which issues and conversations keep AI experts up at night? 1. Unemployment. What happens after the end of jobs? The hierarchy of labour is concerned primarily with automation. As we’ve invented ways to automate jobs, we could create room for people to assume more complex roles, moving from the physical work that dominated the pre-industrial globe to the cognitive labour that characterizes strategic and administrative work in our globalized society. Look at trucking: it currently employs millions of individuals in the United States alone. What will happen to them if the self-driving trucks promised by Tesla’s Elon Musk become widely available in the next decade? But on the other hand, if we consider the lower risk of accidents, self-driving trucks seem like an ethical choice. The same scenario could happen to office workers, as well as to the majority of the workforce in developed countries. Have you read? Artificial Intelligence Collides with Patent Law Robot inventors are on the rise. But are they welcomed by the patent system? Artificial intelligence could be our saviour, according to the CEO of Google This is where we come to the question of how we are going to spend our time. Most people still rely on selling their time to have enough income to sustain themselves and their families. We can only hope that this opportunity will enable people to find meaning in non-labour activities, such as caring for their families, engaging with their communities and learning new ways to contribute to human society. If we succeed with the transition, one day we might look back and think that it was barbaric that human beings were required to sell the majority of their waking time just to be able to live. 2. Inequality. How do we distribute the wealth created by machines? Our economic system is based on compensation for contribution to the economy, often assessed using an hourly wage. The majority of companies are still dependent on hourly work when it comes to products and services. But by using artificial intelligence, a company can drastically cut down on relying on the human workforce, and this means that revenues will go to fewer people. Consequently, individuals who have ownership in AI-driven companies will make all the money. We are already seeing a widening wealth gap, where start-up founders take home a large portion of the economic surplus they create. In 2014, roughly the same revenues were generated by the three biggest companies in Detroit and the three biggest companies in Silicon Valley ... only in Silicon Valley there were 10 times fewer employees. If we’re truly imagining a post-work society, how do we structure a fair post-labour economy? 3. Humanity. How do machines affect our behaviour and interaction? Artificially intelligent bots are becoming better and better at modelling human conversation and relationships. In 2015, a bot named Eugene Goostman won the Turing Challenge for the first time. In this challenge, human raters used text input to chat with an unknown entity, then guessed whether they had been chatting with a human or a machine. Eugene Goostman fooled more than half of the human raters into thinking they had been talking to a human being. This milestone is only the start of an age where we will frequently interact with machines as if they are humans; whether in customer service or sales. While humans are limited in the attention and kindness that they can expend on another person, artificial bots can channel virtually unlimited resources into building relationships. Even though not many of us are aware of this, we are already witnesses to how machines can trigger the reward centres in the human brain. Just look at click-bait headlines and video games. These headlines are often optimized with A/B testing, a rudimentary form of algorithmic optimization for content to capture our attention. This and other methods are used to make numerous video and mobile games become addictive. Tech addiction is the new frontier of human dependency. On the other hand, maybe we can think of a different use for software, which has already become effective at directing human attention and triggering certain actions. When used right, this could evolve into an opportunity to nudge society towards more beneficial behavior. However, in the wrong hands it could prove detrimental. 4. Artificial stupidity. How can we guard against mistakes? Intelligence comes from learning, whether you’re human or machine. Systems usually have a training phase in which they "learn" to detect the right patterns and act according to their input. Once a system is fully trained, it can then go into test phase, where it is hit with more examples and we see how it performs. Obviously, the training phase cannot cover all possible examples that a system may deal with in the real world. These systems can be fooled in ways that humans wouldn't be. For example, random dot patterns can lead a machine to “see” things that aren’t there. If we rely on AI to bring us into a new world of labour, security and efficiency, we need to ensure that the machine performs as planned, and that people can’t overpower it to use it for their own ends.

AI Researchers are beginning to recognize biases and resolve them---means the squo will solve the links

Berreby 20 [David Berreby, 11-22-2020, accessed on 6-25-2022, The New York Times, "Can We Make Our Robots Less Biased Than We Are?", <https://www.nytimes.com/2020/11/22/science/artificial-intelligence-robots-racism-police.html>] -os-

On a summer night in Dallas in 2016, a bomb-handling robot made technological history. Police officers had attached roughly a pound of C-4 explosive to it, steered the device up to a wall near an active shooter and detonated the charge. In the explosion, the assailant, Micah Xavier Johnson, became the first person in the United States to be killed by a police robot. Afterward, then-Dallas Police Chief David Brown called the decision sound. Before the robot attacked, Mr. Johnson had shot five officers dead, wounded nine others and hit two civilians, and negotiations had stalled. Sending the machine was safer than sending in human officers, Mr. Brown said. But some robotics researchers were troubled. “Bomb squad” robots are marketed as tools for safely disposing of bombs, not for delivering them to targets. (In 2018, police officers in Dixmont, Maine, ended a shootout in a similar manner.). Their profession had supplied the police with a new form of lethal weapon, and in its first use as such, it had killed a Black man. “A key facet of the case is the man happened to be African-American,” Ayanna Howard, a robotics researcher at Georgia Tech, and Jason Borenstein, a colleague in the university’s school of public policy, wrote in a 2017 paper titled “The Ugly Truth About Ourselves and Our Robot Creations” in the journal Science and Engineering Ethics. Like almost all police robots in use today, the Dallas device was a straightforward remote-control platform. But more sophisticated robots are being developed in labs around the world, and they will use artificial intelligence to do much more. A robot with algorithms for, say, facial recognition, or predicting people’s actions, or deciding on its own to fire “nonlethal” projectiles is a robot that many researchers find problematic. The reason: Many of today’s algorithms are biased against people of color and others who are unlike the white, male, affluent and able-bodied designers of most computer and robot systems. While Mr. Johnson’s death resulted from a human decision, in the future such a decision might be made by a robot — one created by humans, with their flaws in judgment baked in. “Given the current tensions arising from police shootings of African-American men from Ferguson to Baton Rouge,” Dr. Howard, a leader of the organization Black in Robotics, and Dr. Borenstein wrote, “it is disconcerting that robot peacekeepers, including police and military robots, will, at some point, be given increased freedom to decide whether to take a human life, especially if problems related to bias have not been resolved.” Last summer, hundreds of A.I. and robotics researchers signed statements committing themselves to changing the way their fields work. One statement, from the organization Black in Computing, sounded an alarm that “the technologies we help create to benefit society are also disrupting Black communities through the proliferation of racial profiling.” Another manifesto, “No Justice, No Robots,” commits its signers to refusing to work with or for law enforcement agencies. Over the past decade, evidence has accumulated that “bias is the original sin of A.I,” Dr. Howard notes in her 2020 audiobook, “Sex, Race and Robots.” Facial-recognition systems have been shown to be more accurate in identifying white faces than those of other people. (In January, one such system told the Detroit police that it had matched photos of a suspected thief with the driver’s license photo of Robert Julian-Borchak Williams, a Black man with no connection to the crime.) There are A.I. systems enabling self-driving cars to detect pedestrians — last year Benjamin Wilson of Georgia Tech and his colleagues found that eight such systems were worse at recognizing people with darker skin tones than paler ones. Joy Buolamwini, the founder of the Algorithmic Justice League and a graduate researcher at the M.I.T. Media Lab, has encountered interactive robots at two different laboratories that failed to detect her. (For her work with such a robot at M.I.T., she wore a white mask in order to be seen.) The long-term solution for such lapses is “having more folks that look like the United States population at the table when technology is designed,” said Chris S. Crawford, a professor at the University of Alabama who works on direct brain-to-robot controls. Algorithms trained mostly on white male faces (by mostly white male developers who don’t notice the absence of other kinds of people in the process) are better at recognizing white males than other people. “I personally was in Silicon Valley when some of these technologies were being developed,” he said. More than once, he added, “I would sit down and they would test it on me, and it wouldn’t work. And I was like, You know why it’s not working, right?” Robot researchers are typically educated to solve difficult technical problems, not to consider societal questions about who gets to make robots or how the machines affect society. So it was striking that many roboticists signed statements declaring themselves responsible for addressing injustices in the lab and outside it. They committed themselves to actions aimed at making the creation and usage of robots less unjust.

## Perm

### Do both

#### Perm do both – the pitfalls of AI are best solved by political solutions by nation states that utilize existing technologies to create legally binding regulatory policies to combat overambitions

**Sharkey 18** (Noel, Emeritus Professor at the University of Sheffield, 8-28-2018, accessed on 6-25-2022, Humanitarian Law & Policy Blog, "The impact of gender and race bias in AI - Humanitarian Law & Policy Blog", https://blogs.icrc.org/law-and-policy/2018/08/28/impact-gender-race-bias-ai/)-qcl

It should be clear from evidence presented above that both AI decision algorithms and face recognition algorithms can be alarmingly biased or inaccurate with darker shades of skin and with women. These may well improve over time but there have been no magic bullet solutions despite massive efforts and several announcements. Many of the companies developing software, particularly for policing, insist that they did well on their inhouse testing. It has remained for other organisations, such as NGOs, to collect the data and demonstrate the biases, yet the systems keep on getting rolled out. It is the familiar old story that once there has been huge investment in a technology it continues to be used despite its failings. Let us not make the same mistake with targeting technology. Discriminatory systems are bad enough in the civilian world where new cases of injustice to women and people with darker shades of skin are turning up almost weekly. But while it can be difficult for those who suspect discrimination to take legal action, there is at least the potential to reverse such unjust decisions. It is a different story when dealing with the technologies of violence. Once someone has been misclassified and targeted with lethal force by an unfairly biased decision process, there is no overturning the decision. Technology, and particularly AI, has always gotten ahead of itself with ambition outstripping achievement. In my long experience working on the subject and reviewing many research proposals, ambition often wins the day. Indeed, ambition is often a positive step towards achievement. In many cases it can still be worthwhile even if the achievement falls well short of the ambition. However, when it comes to technologies of violence, we need to be considerably more cautious of ambitious claims about speculative technology that can lead us down the wrong path. Like a retired police horse, it is time to take off the blinkers and look at the current state of technology and its problematic relationship to the technologies of violence. We cannot simply ignore the types of discriminatory algorithmic biases appearing in the civilian world and pretend that we can just make them go away when it comes weapons development and use. These are just some of the problems that have come to light, since the increased use of AI in society. We don’t know what further problems are around the corner or what further biases are likely to occur in targeting technologies. The moral of this tale is simple. We must take a precautionary approach to the use of AI in weapons technology and AWS in particular. We must not rely on the possibility of future fixes but instead make decisions based on what the technology is capable of today. It is time now for nation States to step up to the mark and begin negotiations for a new international legally binding instrument to ensure the meaningful human control of weapons systems is preserved.

### Bias

#### Perm We will never get perfect bias mitigation – but most effective approach is all-encompassing

UNIDIR (United Nations Institute for Disarmament Research), 2018, “Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies – A Primer”, The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to a variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and Governments. UNIDIR activities are funded by contributions from Governments and donor foundations, <https://www.unidir.org/sites/default/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf>, - Maren Lien

As algorithms approach ubiquity, there is growing understanding that they are not objective and infallible. Algorithms in **all** domains, including military applications, can exhibit multiple types of **biases** that arise from **different sources**, such as unrepresentative training data or inappropriate transfer of the algorithm to a novel context. Some degree of **algorithmic bias** may be **inevitable**, as it might not be possible to satisfy all relevant norms with a single process, decision, or algorithm. At the same time, algorithmic biases are not mutually exclusive, as some biases feed into one another. Moreover, not all biases are bad, as some biases can be beneficial to achieving the user’s end goals. Most pointedly, **algorithmic** **bias** can arise at every stage of **development** and **deployment**, with each stage bringing its own set of considerations and possibilities for the outcome of bias. In many cases, **mitigation** strategies are **available**, but they require **careful** engagement with the details of the **situation**, as one might not want to mitigate; or might be able to mitigate only some biases; or might address problems by changing the users or broader system; and so forth. Various institutions and organizations are beginning to address these challenges, though policy and technical responses are still in their infancy. As a contribution to the policy response, those participating in the discussion on LAWS within the CCW framework may wish to consider the following questions about algorithmic biases in future systems: • If governments decide to regulate increasingly autonomous weapon systems, rather than adopt an outright ban, which national or international organizations or instruments would be best placed to offer guidance or assistance to address potential algorithmic biases in AWS, including identifying possible mitigation steps? • Given the secretive or non-transparent nature of weapon development and weapon review processes, what sorts of “best practices” can provide confidence that key algorithmic biases have been appropriately identified and mitigated? • Are mitigation steps for algorithmic biases in particular AWS robust against possible loss of communication, interoperability challenges, or reduced human oversight? • How would **training** of **operators** and **commanders** need to be **adapted** to ensure that they appropriately understand the **algorithmic biases** in an **AWS**, in order to maintain **trust** in the **system** and ensure its **lawful** use?

### US leadership

#### Perm - US is most favorable to ethical principles for AI weapons

David H. Freedman, 9/15/21, “US Is Only Nation with Ethical Standards for AI Weapons. Should We Be Afraid?”, Newsweek, David H. Freedman is a scientific journalist, author, and is a contributing writer at The Atlantic and Newsweek, <https://www.newsweek.com/2021/09/24/us-only-nation-ethical-standards-ai-weapons-should-we-afraid-1628986.html> - Maren Lien

Even if military AI systems work exactly as intended, is it ethical to give machines the authority to destroy and kill? Work, the former defense deputy secretary, insists the **U.S.** **military** is strictly committed to keeping a **human** **decision-maker** in the "**kill chain**" so that no **weapon** will pick a target and **fire on its own** without an OK. But other nations may not be as careful, he says. "As far as we know, the **U.S. military** is the **only one** that has established **ethical** principles for **AI." Twenty-two nations** have asked the **U**nited **N**ations to **ban** automated weapons capable of operating **outside human oversight**, but so far **no agreements** have been signed. Human Rights Watch and other advocacy groups have called for similar bans to no avail.

#### Perm True mitigation requires multiple sets of actors – including international government regulators

UNIDIR (United Nations Institute for Disarmament Research), 2018, “Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies – A Primer”, The United Nations Institute for Disarmament Research (UNIDIR)—an autonomous institute within the United Nations—conducts research on disarmament and security. UNIDIR is based in Geneva, Switzerland, the centre for bilateral and multilateral disarmament and non-proliferation negotiations, and home of the Conference on Disarmament. The Institute explores current issues pertaining to a variety of existing and future armaments, as well as global diplomacy and local tensions and conflicts. Working with researchers, diplomats, government officials, NGOs and other institutions since 1980, UNIDIR acts as a bridge between the research community and Governments. UNIDIR activities are funded by contributions from Governments and donor foundations, <https://www.unidir.org/sites/default/files/publication/pdfs/algorithmic-bias-and-the-weaponization-of-increasingly-autonomous-technologies-en-720.pdf>, - Maren Lien

Responsibility for **mitigating** unwanted **algorithmic biases** does **not** rest with a **single actor**. A first set of **actors** are the **program developers** designing and creating the system. The developer is intimately familiar with each of the algorithms running in the system. To the extent that an undesirable bias can be mitigated through changes in the underlying algorithms or development process, then developers present a natural locus of intervention. In this way, some potential problems can be avoided before the system is fully built. At the same time, not all algorithmic biases can be addressed purely in the development stage. For example, appropriate training data might not be available, and the developers might have insufficient knowledge of deployment contexts to appropriately adjust their algorithms. The second set of key actors in potential **mitigation** of AWS algorithm biases are the **acquirers** of the technology. The agency or organization responsible for the purchase of the technology can require that the system have certain features, or meet specific, pre-defined standards. Alternately, the acquirer can require that the developers provide them with precise, detailed information about the training data, intended use contexts, and so forth. In the former case, the acquirer indicates which algorithmic biases are unacceptable, and the developer must find some way of producing such a system. In the latter case, the acquirer gains the knowledge needed to adapt practices (such as rules of engagement) to minimize the harms from the algorithmic biases that remain. In either case, acquisition and procurement teams can minimize the likelihood of algorithmic “failures” or negative biases. The third set of potential actors in **mitigation** efforts are **regulators** (including **international policymakers**) and testers. Regulators could decide to **completely ban** the **development** or **use** of **AWS**. Alternatively, they may decide to restrict or regulate some facet of development or use. In this case, they may determine which algorithmic **biases** are **unacceptable**, and not allow **deployment** of systems that **exhibit** those **biases**. They could prioritize various conditions, properties, and behaviours of a weapon system, and thereby impose particular **ethical**, **legal**, or **social norms** that the **system** must **follow**, though the developers are left with the task of determining how to satisfy those constraints. **National** or **international regulators** also have the ability to **dictate regulatory constraints and processes** that can help guide developers and future testers in their search for these or similar-acting system biases. Lastly, through testing, some algorithmic biases may be identified prior to approval and deployment, allowing for system revisions prior to the negative, real-world or real-life impacts that would impair efficacy or trust in future AWS deployment. The fourth set of potential actors would be the deployers or operators of the system. These actors, whether at the strategic or tactical level, would make the final decisions about whether, when and where to use the weapon system, and so have the ability to mitigate algorithmic biases simply by not using the system. Alternately, if a system is used only in settings for which it was designed with appropriate training data (and all of the other conditions), then the system’s potentially harmful impacts will be mitigated—though not necessarily completely eliminated.

### Governance

#### AI can be made ethical through policy and collaboration

Blackman 20 Harvard Business Review, 10-15-2020, "A Practical Guide to Building Ethical AI," <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai> [AJL]

How to Operationalize Data and AI Ethics AI ethics does not come in a box. Given the varying values of companies across dozens of industries, a data and AI ethics program must be tailored to the specific business and regulatory needs that are relevant to the company. However, here are seven steps towards building a customized, operationalized, scalable, and sustainable data and AI ethics program. 1. Identify existing infrastructure that a data and AI ethics program can leverage. The key to a successful creation of a data and AI ethics program is using the power and authority of existing infrastructure, such as a data governance board that convenes to discuss privacy, cyber, compliance, and other data-related risks. This allows concerns from those “on the ground” (e.g., product owners and managers) to bubble up and, when necessary, they can in turn elevate key concerns to relevant executives. Governance board buy in works for a few reasons: 1) the executive level sets the tone for how seriously employees will take these issues, 2) a data and AI ethics strategy needs to dovetail with the general data and AI strategy, which is devised at the executive level, and 3) protecting the brand from reputational, regulatory, and legal risk is ultimately a C-suite responsibility, and they need to be alerted when high stakes issues arise. If such a body does not exist then companies can create one — an ethics council or committee, for example — with ethics-adjacent personnel, such as those in cyber, risk and compliance, privacy, and analytics. It may also be advisable to include external subject matter experts, including ethicists. 2. Create a data and AI ethical risk framework that is tailored to your industry. A good framework comprises, at a minimum, an articulation of the ethical standards — including the ethical nightmares — of the company, an identification of the relevant external and internal stakeholders, a recommended governance structure, and an articulation of how that structure will be maintained in the face of changing personnel and circumstances. It is important to establish KPIs and a quality assurance program to measure the continued effectiveness of the tactics carrying out your strategy. A robust framework also makes clear how ethical risk mitigation is built into operations. For instance, it should identify the ethical standards data collectors, product developers, and product managers and owners must adhere to. It should also articulate a clear process by which ethical concerns are elevated to more senior leadership or to an ethics committee. All companies should ask whether there are processes in place that vet for biased algorithms, privacy violations, and unexplainable outputs. Still, frameworks also need to be tailored to a company’s industry. In finance, it is important to think about how digital identities are determined and how international transactions can be ethically safe. In health care there will need to be extra protections built around privacy, particularly as AI enables the development of precision medicine. In the retail space, where recommendation engines loom large, it is important to develop methods to detect and mitigate associative bias, where recommendations flow from stereotypical and sometimes offensive associations with various populations. 3. Change how you think about ethics by taking cues from the successes in health care. Many senior leaders describe ethics in general — and data and AI ethics in particular — as “squishy” or “fuzzy,” and argue it is not sufficiently “concrete” to be actionable. Leaders should take inspiration from health care, an industry that has been systematically focused on ethical risk mitigation since at least the 1970s. Key concerns about what constitutes privacy, self-determination, and informed consent, for example, have been explored deeply by medical ethicists, health care practitioners, regulators, and lawyers. Those insights can be transferred to many ethical dilemmas around consumer data privacy and control. For instance, companies attest to respect the users of their products, but what does that mean in practice? In health care, an essential requirement of demonstrating respect for patients is that they are treated only after granting their informed consent — understood to include consent that, at a minimum, does not result from lies, manipulation, or communications in words the patient cannot understand, such as impenetrable legalese or Latin medical terms. These same kinds of requirements can be brought to bear on how people’s data is collected, used, and shared. Ensuring that users are not only informed of how their data is being used, but also that they are informed early on and in a way that makes comprehension likely (for instance, by not burying the information in a long legal document), is one easy lesson to take from health care. The more general lesson is to break down big ethical concepts like privacy, bias, and explainability into infrastructure, process, and practice that realize those values. 4. Optimize guidance and tools for product managers. While your framework provides high-level guidance, it’s essential that guidance at the product level is granular. Take, for instance, the oft-lauded value of explainability in AI, a highly valued feature of ML models that will likely be part of your framework. Standard machine-learning algorithms engage in pattern recognition too unwieldy for humans to grasp. But it is common — particularly when the outputs of the AI are potentially life-altering — to want or demand explanations for AI outputs. The problem is that there is often a tension between making outputs explainable, on the one hand, and making the outputs (e.g. predictions) accurate, on the other. Product managers need to know how to make that tradeoff, and customized tools should be developed to help product managers make those decisions. For example, companies can create a tool by which project managers can evaluate the importance of explainability for a given product. If explainability is desirable because it helps to ferret out bias in an algorithm, but biased outputs are not a concern for this particular ML application, then that downgrades the importance of explainability relative to accuracy. If the outputs fall under regulations that require explanations — for instance, regulations in the banking industry that require banks to explain why someone has been turned down for a loan — then explainability will be imperative. The same goes for other relevant values, e.g. which, if any, of the dozens of metrics to use when determining whether a product delivers fair or equitable outputs. 5. Build organizational awareness. Ten years ago, corporations scarcely paid attention to cyber risks, but they certainly do now, and employees are expected to have a grasp of some of those risks. Anyone who touches data or AI products — be they in HR, marketing, or operations — should understand the company’s data and AI ethics framework. Creating a culture in which a data and AI ethics strategy can be successfully deployed and maintained requires educating and upskilling employees, and empowering them to raise important questions at crucial junctures and raise key concerns to the appropriate deliberative body. Throughout this process, it’s important to clearly articulate why data and AI ethics matters to the organization in a way that demonstrates the commitment is not merely part of a public relations campaign. 6. Formally and informally incentivize employees to play a role in identifying AI ethical risks. As we’ve learned from numerous infamous examples, ethical standards are compromised when people are financially incentivized to act unethically. Similarly, failing to financially incentivize ethical actions can lead to them being deprioritized. A company’s values are partly determined by how it directs financial resources. When employees don’t see a budget behind scaling and maintaining a strong data and AI ethics program, they will turn their attention to what moves them forward in their career. Rewarding people for their efforts in promoting a data ethics program is essential. 7. Monitor impacts and engage stakeholders. Creating organizational awareness, ethics committees, informed product managers owners, engineers, and data collectors is all part of the development and, ideally, procurement process. But due to limited resources, time, and a general failure to imagine all the ways things can go wrong, it is important to monitor the impacts of the data and AI products that are on the market. A car can be built with air bags and crumple zones, but that doesn’t mean it’s safe to drive it at 100 mph down a side street. Similarly, AI products can be ethically developed but unethically deployed. There is both qualitative and quantitative research to be done here, including especially engaging stakeholders to determine how the product has affected them. Indeed, in the ideal scenario, relevant stakeholders are identified early in the development process and incorporated into an articulation of what the product does and does not do. Operationalizing data and AI ethics is not an easy task. It requires buy-in from senior leadership and cross-functional collaboration. Companies that make the investment, however, will not only see mitigated risk but also more efficient adoption of the technologies they need to forge ahead. And finally, they’ll be exactly what their clients, consumers, and employees are looking for: trustworthy.

### Distributed cognition

#### **Distributed cognitions models enhance human life – post humanist ideologies help us fashion new modes of existence and explore the potential of virtual technologies**

Hayles ’99 (N. Katherine Hayles is a literary critic and theorist. She is the author of *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature and Informatics* which won the Rene Wellek Prize for the best book in literary theory for 1998–1999.--“How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics”—Feb 15, 1999—book--<https://monoskop.org/images/5/50/Hayles_N_Katherine_How_We_Became_Posthuman_Virtual_Bodies_in_Cybernetics_Literature_and_Informatics.pdf)//Marzz>

Hutchins would no doubt disagree with Weizenbaum's view that judgment should be reserved for humans alone. Like cognition, decision making is distributed between human and nonhuman agents, from the steam-powered steering system that suddenly failed on a navy vessel Hutchins was studying to the charts and pocket calculators that the navigators were then forced to use to calculate their position. He convincingly shows that these adaptations to changed circumstances were evolutionary and embodied rather than abstract and consciously designed (pp. 347-51). The solution to the problem caused by this sudden failure of the steering mechanism was" clearly discovered by the organization [of the system as a whole] before it was discovered by any of the participants" (p. 361). Seen in this perspective, the prospect of humans working in partnership with intelligent machines is not so much a usurpation of human right and responsibility as it is a further development in the construction of distributed cognition environments, a construction that has been ongoing for thousands of years. Also changed in this perspective is the relation of human subjectivity to its environment. No longer is human will seen as the source from which emanates the mastery necessary to dominate and control the environment. Rather, the distributed cognition of the emergent human subject correlates with-in Bateson's phrase, becomes a metaphor for-the distributed cognitive system as a whole, in which "thinking" is done by both human and nonhuman actors. "Thinking consists of bringing these structures into coordination so they can shape and be shaped by one another," Hutchins wrote (p. 316). To conceptualize the human in these terms is not to imperil human survival but is precisely to enhance it, for the more we understand the flexible, adaptive structures that coordinate our environments and the metaphors that we ourselves are, the better we can fashion images of ourselves that accurately reflect the **complex interplays** that ultimately make the entire world one system. This view of the posthuman also offers resources for thinking in more sophisticated ways about virtual technologies. As long as the human subject is envisioned as an autonomous self with unambiguous boundaries, the human-computer interface can only be parsed as a division between the solidity of real life on one side and the illusion of virtual reality on the other, thus obscuring the far-reaching changes initiated by the development of virtual technologies. Only if one thinks of the subject as an autonomous self independent of the environment is one likely to experience the panic performed by Norbert Wiener's Cybernetics and Bernard Wolfe's Limbo. This view of the self authorizes the fear that if the boundaries are breached at all, there will be nothing to stop the self's complete dissolution. By contrast, when the human is seen as part of a distributed system, the full expression of human capability can be seen precisely to depend on the splice rather than being imperiled by it. Writing in another context, Hutchins arrives at an in Sight profoundly applicable to virtual technologies: "What used to look like internalization [of thought and subjectivity] now appears as a gradual propagation of organized functional properties across a set of malleable media" (p. 312). This vision is a potent antidote to the view that parses virtuality as a division between an inert body that is left behind and a disembodied subjectivity that inhabits a virtual realm, the construction of virtuality performed by Case in William Gibson's Neuromancer when he delights in the "bodiless exultation of cyberspace" and fears, above all, dropping back into the "meat" of the body.22 By contrast, in the model that Hutchins presents and that the posthuman helps to authorize, human functionality expands because the parameters of the cognitive system it inhabits expand. In this model, it is not a question of leaving the body behind but rather of extending embodied awareness in highly specific, local, and material ways that would be impossible without electronic prosthesis.

### LAWS

#### LAWS produce ethical and logistic quandaries – weaponizing AI accelerates causalities and human rights abuses

Amnesty International 15 (Amnesty International is a global movement of more than 3 million supporters, members and activists in more than 150 countries and territories who campaign to end grave abuses of human rights. Our vision is for every person to enjoy all the rights enshrined in the Universal Declaration of Human Rights and other international human rights standards. We are independent of any government, political ideology, economic interest or religion and are funded mainly by our membership and public donations.) “AUTONOMOUS WEAPONS SYSTEMS: FIVE KEY HUMAN RIGHTS ISSUES FOR CONSIDERATION” April 10 2015 <https://www.amnesty.org/en/documents/act30/1401/2015/en/> // ZX

Over the past decade, there have been extensive advances in artificial intelligence and other technologies. These will make possible the development and deployment of fully autonomous weapons systems which, once activated, can select, attack, kill and wound human targets, and will be able to operate without effective human control. These weapons systems are often referred to as Lethal Autonomous Robotics (LARs), Lethal Autonomous Weapons Systems (LAWS) and, more comprehensively, Autonomous Weapons Systems (AWS). The rapid development of these weapons systems could not only change the entire nature of warfare, it could also dramatically alter the conduct of law enforcement operations and raises extremely serious human rights concerns, undermining the right to life, the prohibition of torture and other ill-treatment, and the right to security of person, and other human rights. Amnesty International has taken the view that AWS is a useful term for these weapons systems, since these systems can (i) be designed to have lethal or less lethal effects and (ii) be used in armed conflict and/or law enforcement situations. With proliferation they are likely to come to be used by non-state armed groups, criminal gangs and private companies and individuals. Amnesty International takes the term ‘autonomous’ to mean weapons capable of selecting targets and triggering an attack without effective or meaningful human control1 that can ensure the lawful use of force. Such systems would use violence (including less-lethal force) against individuals, and could have adverse consequences for a person’s human rights. While the development of AWS clearly raises serious and legitimate ethical and societal concerns, this briefing paper will examine the implications of AWS in the context of international law, particularly international human rights law and standards. The important concerns around their use in situations of armed conflict, and thus their ability to comply fully with international humanitarian law (IHL), has been the focus of previous work on AWS, including by Human Rights Watch, other members of the Campaign to Stop Killer Robots and the International Committee of the Red Cross (ICRC). This briefing paper, however, will address some of the implications for human rights related to AWS, particularly those rights and standards that govern the conduct of law enforcement operations. Amnesty International believes that the questions surrounding the development and potential use of AWS outside armed conflict (and the ability of such systems to comply with human rights law) are at least as daunting as those related to their use on the battlefield and urgently require attention and consideration2 , ultimately leading to concrete steps that will address this important area of international law. Amnesty International has identified five key human rights issues for consideration in the current debate on AWS: 1) The scope of the Convention on Certain Conventional Weapons (CCW) does not cover non-conflict situations; 2) AWS will not be able to comply with relevant international human rights law (IHRL) and policing standards; 3) Developments in existing semi-autonomous weapons technology pose fundamental challenges for the IHRL framework; 4) In the absence of a prohibition, AWS must be subject to independent weapons reviews; and 5) AWS will erode accountability mechanisms. The issues identified are by no means exhaustive, but rather seek to elucidate the principal concerns around the potential use of AWS in law enforcement operations. This briefing argues that the use of AWS, including less-lethal robotic weapons, in law enforcement operations would be fundamentally incompatible with international human rights law, and would lead to unlawful killings, injuries and other violations of human rights. Furthermore, the use of AWS would pose serious challenges in holding accountable those responsible for serious violations and could entrench impunity for crimes under international law. Consequently, Amnesty International supports the call for a pre-emptive ban on the development, transfer, deployment and use of AWS, including fully autonomous systems that deploy less-lethal weapons and can result in death or serious injury. In the absence of a prohibition, Amnesty International supports the call of UN Special Rapporteur on extrajudicial, summary or arbitrary executions, Christof Heyns, to impose a moratorium on the development, transfer, deployment and use of AWS and ensure that moratorium covers both lethal and less-lethal weapons. This principle deals with two different thresholds: a) when it is appropriate to use firearms (potentially lethal force) and b) the even higher threshold of when the intentional lethal use of firearms is permissible. Each of these situations involves a complex assessment of potential or imminent threats to life or serious injury and how to respond to them appropriately, and it involves deciding how best to protect the right to life, which is an absolutely fundamental duty of the state under human rights law. Such life and death decisions must never be delegated to AWS. In order to be able to carry out policing and law enforcement operations in a lawful manner, AWS would need to be able to effectively assess the degree to which there was an imminent threat of death or serious injury, identify correctly who is posing the threat, consider whether force is necessary to neutralize the threat, be able to identify and use means other than force, have the capacity to deploy different modes of communication and policing weapons and equipment to allow for a graduated response, and have available back up means and resources. To add to this complexity, each situation would require a different and unique response, which would be extremely challenging to reduce to a series of complex algorithms. It is not possible that AWS, without meaningful and effective human control and judgement, would be able to comply with these provisions, especially in unpredictable and ever-evolving environments. In an open letter in October 2013, computer scientists, engineers, artificial intelligence experts, roboticists and professionals from related disciplines from 37 countries asserted that “in the absence of clear scientific evidence that robot weapons have, or are likely to have in the foreseeable future, the functionality required for accurate target identification, situational awareness or decisions regarding the proportional use of force, we question whether they could meet the strict legal requirements for the use of force” and that “[G]iven the limitations and unknown future risks of autonomous robot weapons technology…,[D]ecisions about the application of violent force must not be delegated to machines.”15 The UNBPUFF places a due diligence requirement upon states to review weapons used in law enforcement. As Principle 3 of the UNBPUFF states, “the development and deployment of non-lethal incapacitating weapons should be carefully evaluated in order to minimize the risk of endangering uninvolved persons”. This review is limited to less-lethal weapons but is still important to ensure that those weapons will comply with relevant international standards and national laws and, moreover, given that evidence shows that “non-lethal” weapons can often have lethal effects which is why the term “less-lethal” is more appropriate. The requirement of a review of weapons used for law enforcement is even more important given the increasing ‘militarization’ of law enforcement operations, whereby military personnel assume roles often held by law enforcement agencies, such as policing of public assemblies. In the absence of a prohibition on AWS, states intending to develop, acquire, or use AWS must therefore be required to thoroughly review whether they can be used in a manner that fully respects relevant law and standards be it for law enforcement or military operations. This testing should be carried out by an independent body. The rapid technological advances that are moving towards full autonomy in weapons systems present serious concerns. The technology to allow fully autonomous operations may be reached soon; but it is extremely unlikely that programming that could ensure AWS perform law enforcement functions lawfully would be developed in the foreseeable future. Any new law enforcement equipment should be introduced based on clearly defined operational needs and technical requirements with a view to reduce the amount of force used and the risk and level of harm and injury caused. They must be subject to rigorous testing, by an independent expert body, and the testing, review and selection process should be legally constituted. In addition to assessing compliance with the UNBPUFF themselves, the process must test AWS compatibility with other key human rights treaties and standards, including ICCPR, International Covenant on Economic, Social and Cultural Rights (CESCR), the Convention Against Torture, the SMRTP and the UNCCLEO.

## AI Turns

#### Companies use AI in a way that has a significant impact on people’s lives

Kearns and Roth 19 Kearns, Michael, and Aaron Roth. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Illustrated, Oxford University Press, 2019. [AJL]

Which all brings us to a conundrum. The insights we can get from this unprecedented access to data can be a great thing: we can get new understanding about how our society works, and improve public health, municipal services, and consumer products. But as individuals, we aren’t just the recipients of the fruits of this data analysis: we are the data, and it is being used to make decisions about us—sometimes very consequential decisions. In December 2018, the New York Times obtained a commercial dataset containing location information collected from phone apps whose nominal purpose is to provide mundane things like weather reports and restaurant recommendations. Such datasets contain precise locations for hundreds of millions of individuals, each updated hundreds (or even thousands) of times a day. Commercial buyers of such data will generally be interested in aggregate information—for example, a hedge fund might be interested in tracking the number of people who shop at a particular chain of retail outlets in order to predict quarterly revenues. But the data is recorded by individual phones. It is superficially anonymous, without names attached—but there is only so much anonymity you can promise when recording a person’s every move. For example, from this data the New York Times was able to identify a forty-six-year-old math teacher named Lisa Magrin. She was the only person who made the daily commute from her home in upstate New York to the middle school where she works, fourteen miles away. And once someone’s identity is uncovered in this way, it’s possible to learn a lot more about them. The Times followed Lisa’s data trail to Weight Watchers, to a dermatologist’s office, and to her ex-boyfriend’s home. She found this disturbing and told the Times why: “It’s the thought of people finding out those intimate details that you don’t want people to know.” Just a couple of decades ago, this level of intrusive surveillance would have required a private investigator or a Introduction ■ 3 government agency; now it is simply the by-product of widely available commercial datasets. Clearly, we have entered a brave new world. And it’s not only privacy that has become a concern as data gathering and analysis proliferate. Because algorithms—those little bits of machine code that increasingly mediate our behavior via our phones and the Internet—aren’t simply analyzing the data that we generate with our every move. They are also being used to actively make decisions that affect our lives. When you apply for a credit card, your application may never be examined by a human being. Instead, an algorithm pulling in data about you (and perhaps also about people “like you”) from many different sources might automatically approve or deny your request. Though there are benefits to knowing instantaneously whether your request is approved, rather than waiting five to ten business days, this should give us a moment of pause. In many states, algorithms based on what is called machine learning are also used to inform bail, parole, and criminal sentencing decisions. Algorithms are used to deploy police officers across cities. They are being used to make decisions in all sorts of domains that have direct and real impact on people’s lives. All this raises questions not only of privacy but also of fairness, as well as a variety of other basic social values including safety, transparency, accountability, and even morality. So if we are going to continue to generate and use huge datasets to automate important decisions (a trend whose reversal seems about as plausible as our returning to an agrarian society), we have to think seriously about some weighty topics. These include limits on the use of data and algorithms, and the corresponding laws, regulations, and organizations that would determine and enforce those limits. But we must also think seriously about addressing the concerns scientifically—about what it might mean to encode ethical principles directly into the design of the algorithms that are increasingly 4 ■ THE ETHICAL ALGORITHM woven into our daily lives. This book is about the emerging science of ethical algorithm design, which tries to do exactly that.

#### AI can help in international peace efforts – no existential threat from AI

**Daanish** Masood and **Martin** Waehlisch, 2019-04-23, “AI & Global Governance: Robots Will Not Only Wage Future Wars also Future Peace, United Nations University Centre for Policy Research**, Daanish Masood and Martin Waehlisch are Political Affairs Officers at the UN’s Department of Political and Peacebuilding Affairs.** <https://cpr.unu.edu/publications/articles/robots-will-not-only-wage-future-wars-but-also-future-peace.html> **- Maren Lien**

Though touted as a real possibility by the likes of Elon Musk, that particular idea has been dismissed in the field as far-fetched. In his 2018 book, Ten Arguments for Deleting Your Social Media Accounts Now, polymathic computer scientist and ‘founding father’ of virtual reality Jaron Lanier described AI as a decades-old lie that he and others in Silicon Valley invented just to get money from DARPA, the US Pentagon agency responsible for researching technological breakthroughs. Lanier was being tongue-in-cheek. His point was that despite our dystopian fears, **AI** is still far too **rudimentary** to pose an **existential threat** to the **human species**. At the United Nations, we have been exploring completely different scenarios for AI: its **potentia**l to be used for the noble purposes of **peace and security**. This could **revolutionize** the way of how we **prevent** and **solve conflicts globally**. Two of the most promising areas are Machine Learning and Natural Language Processing. Machine Learning involves computer algorithms detecting patterns from data to learn how to make predictions and recommendations. Natural Language Processing involves computers learning to understand human languages. At the UN Secretariat, our chief concern is with how these emerging technologies can be deployed for the good of humanity to de-escalate violence and increase international stability. This endeavor has admirable precedent. During the **Cold War**, computer scientists used **multilayered** simulations to predict the **scale** and **potential** outcome of the arms race between the East and the West. Since then, governments and international agencies have **increasingly** used **computational** **models** and advanced Machine Learning to try to understand recurrent conflict patterns and forecast moments of state fragility. But two things have transformed the scope for progress in this field. The first is the sheer volume of data now available from what people say and do online. The second is the game-changing growth in computational capacity that allows us to crunch unprecedented, inconceivable quantities data with relative speed and ease. So how can this help the United Nations build peace? Three ways come to mind. Firstly, overcoming cultural and language barriers. By teaching **computers** to understand human language and the nuances of dialects, not only can we better link up what people write on social media to local contexts of conflict, we can also more methodically follow what people **say** on radio and TV. As part of the UN’s early warning efforts, this can help us **detect hate speech** in a place where the potential for **conflict** is high. This is crucial because the UN often works in countries where internet coverage is low, and where the spoken languages may not be well understood by many of its international staff. Natural Language Processing algorithms can help to **track** and improve understanding of **local debates**, which might well be **blind spots** for the international community. If we combine such methods with Machine Learning chatbots, the UN could conduct large-scale digital focus groups with thousands in real-time, enabling different demographic segments in a country to voice their views on, say, a proposed peace deal – instantly testing public support, and indicating the chances of sustainability. Secondly, anticipating the deeper drivers of conflict. We could combine new imaging techniques – whether satellites or drones – with automation. For instance, many parts of the world are experiencing severe groundwater withdrawal and water aquifer depletion. Water scarcity, in turn, drives conflicts and undermines stability in post-conflict environments, where violence around water access becomes more likely, along with large movements of people leaving newly arid areas. One of the best predictors of water depletion is land subsidence or sinking, which can be measured by satellite and drone imagery. By combining these **imaging techniques** with **Machine Learning**, the UN can work in **partnership** with **governments and local communities** to **anticipate** future **water conflicts** and begin working **proactively** to **reduce** their likelihood. Thirdly, advancing decision making. In the work of peace and security, it is surprising how many consequential decisions are still made solely on the basis of intuition. Yet **complex** decisions often need to navigate conflicting goals and undiscovered options, against a landscape of **limited information** and political preference. This is where we can use Deep Learning – where a **network** can absorb **huge** amounts of public **data** and **test** it against **real-world** examples on which it is trained while applying with probabilistic modeling. This mathematical approach can help us to generate models of our uncertain, dynamic world with limited data. With better data, we can eventually make better predictions to guide **complex** decisions. Future senior peace envoys charged with mediating a conflict would benefit from such advances to stress test elements of a peace agreement. Of course, **human decision**-making will remain crucial, but would be **informed** by more **evidence-driven** robust **analytical** tools. Doing the above inside the UN, will require training staff and senior leaders in new approaches and trusting in their competence. And it will also require collaborating with university researchers, and forging close partnerships with leading private AI and technology firms. The good news is that the work has already started. But we are still at baby-steps. With the Secretary-General’s support, including through his landmark Strategy on New Technologies, the time to scale this activity has come. We can leave no stone unturned and no tool ignored to reduce violence and promote peace – that, after all, is the moral obligation at the very core of the UN Charter.

## AT Alt

### Ivory tower DA

#### Academic alts fail – they don’t make useable alts

Reid Blackman, October 15 2020, “A Practical Guide to Building Ethical AI”, Harvard Business Review, Reid Blackman, Ph.D., is the author of the book Ethical Machines (Harvard Business Review Press, July 2022) and Founder and CEO of Virtue, an AI ethical risk consultancy. He has also been a Senior Advisor to the Deloitte AI Institute, a Founding Member of Ernst & Young’s AI Advisory Board, and volunteers as the Chief Ethics Officer to the non-profit Government Blockchain Association. Reid’s expertise is relied upon by Fortune 500 and Global 1000 companies to speak to and educate their people and to guide them as they create and scale AI ethical risk programs., <https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai>

First, there is the **academic approach**. **Academics** — and I speak from 15 years of experience as a former professor of philosophy — are fantastic at **rigorous** and **systematic** inquiry. Those academics who are ethicists (typically found in philosophy departments) are adept at spotting ethical problems, their sources, and how to think through them. But while academic ethicists might seem like a perfect match, given the need for systematic identification and mitigation of ethical risks, they unfortunately tend to ask **different** questions than businesses. For the most part, academics ask, “Should we do this? Would it be good for society overall? Does it conduce to human flourishing?” Businesses, on the other hand, tend to ask, “Given that we are going to do this, how can we do it without making ourselves vulnerable to ethical risks?” The result is academic treatments that **do not** speak to the **highly particular**, **concrete** uses of **data and AI**. This translates to the **absence** of **clear directives** to the developers on the ground and the senior leaders who need to identify and choose among a set of **risk mitigation** strategies.