An End-to-End Approach to Ethical AI: Socio-Economic Dimensions of the Production and Deployment of Automated Technologies

Antonio A. Casilli, Paola Tubaro

▶ **To cite this version:**

Antonio A. Casilli, Paola Tubaro. An End-to-End Approach to Ethical AI: Socio-Economic Dimensions of the Production and Deployment of Automated Technologies. 2022. hal-04027470

HAL Id: hal-04027470

https://polytechnique.hal.science/hal-04027470v1

Preprint submitted on 13 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An End-to-End Approach to Ethical AI: Socio-Economic Dimensions of the Production and Deployment of Automated Technologies

Antonio A. Casilli & Paola Tubaro

(Forthcoming in L. Robinson, S. Rogerson, K. Moles (eds.) (2023). *Handbook of Digital Social Science*, Edward Elgar.)

INTRODUCTION

Although the beginnings of the field of study of Artificial Intelligence (AI) date back to the 1950s, it is in response to its most recent advances that efforts to devise ethical guidelines have been mushrooming. The successes of commercial applications of AI in the past decade have been largely driven by a shift of focus on a specific paradigm, based on machine learning and enabled by massive digital data together with unprecedented computing power. Unlike early deductive approaches to AI that required coding sets of rules for all possible occurrences, machine-learning algorithms inductively recognise patterns in vast masses of data and use them as a basis to identify trends and make predictions. From simple spam filters to ubiquitous image recognition tools and large language models, these automated systems can improve their accuracy when exposed to increasingly rich information.

In the wake of the growing scientific, political and industrial interest in these models, attention has been drawn to the capacity of AI systems to classify, to predict, and to autonomously make decisions. Nevertheless, the ways they aim to influence or control behaviours may conflict with human rights. As it happened with past technological novelties, real and potential risks take centre stage in the public debate. Compared to other innovations, from cars and television to nuclear power, a particular reason for concern is AI's capacity to operate at scale, with far-reaching impacts.

Surely, some of the concerns voiced in the public arena are based on unrealistic expectations, such as the assumption that an artificial general intelligence (AGI) is actually attainable. While the theoretical possibility of a super-intelligent AI that reproduces or overcomes human-level intelligence may be a subject of philosophical

speculation, its realisation remains largely science-fictional as of today. Epistemologically, the only existing AI is a 'narrow' one, capable of executing a given task but incapable of extending operations outside its predetermined scope (Fjelland, 2020). The forms of intelligent automation currently in use are rather mundane and immanent to our existence: for example, voice assistants such as Siri, Cortana and Google Now are literally 'in our pockets'.

Other concerns are more tangible or based on already existing evidence. The need of machine learning models to access increasing amounts of data for their development, including data that is private or personally identifiable, raises issues of privacy. Surveillance is no longer exclusively in the hands of governments and police forces, but it also involves machine learning-powered intrusions by businesses and private actors (Tubaro, Casilli and Sarabi, 2014; Casilli, 2015). Further, the very functioning of machine learning, which reaches solutions by extracting previously unseen regularities from a given set of data, challenges transparent and democratic decision-making because even the programmer cannot systematically know which characteristics of the data the system has used (Wachter, Mittelstadt and Floridi, 2017). Bias and discrimination have been widely exposed, showing how machines have 'learned' from imbalanced historical data to recognize light-skinned men more accurately than dark-skinned women (Buolamwini and Gebru, 2018), to select men against women in recruitment processes, or to predict higher re-offending risks for black rather than white defendants (Müller, 2021).

These are just a few examples, and any attempt to draw a comprehensive list of ethical concerns would be ephemeral, as the field of AI ethics is in its infancy and grows fast. It is useful, though, to discuss the state of the field itself and how its development has been at the same time fuelled and constrained by the commercial interests, global power imbalances, and labour-capital relations that are shaping the features and usages of the technology. We argue that interpreting AI in a social-scientific perspective that highlights its industrial production structures offers insights that escape both the disembodied views that inform most of today's public debates, and the narrow technical approaches of many engineers and computing specialists. It is a way to bring to light major limitations of extant AI ethics discourses and to propose social science-informed pathways to overcome them. Our approach highlights an important aspect of the relationship between digital technology and the social sciences, noting how the former may gain from the latter's capacity to identify, investigate, and ideally reshape the social, economic, and political factors that get built into tech products.

RECENT PROLIFERATION OF AI ETHICS GUIDELINES: TECHNICAL SOLUTIONS FOR SOCIO-ECONOMIC PROBLEMS?

AI ethics was little heard of in the early 2010s. But a 2015 open letter 'for robust and beneficial Artificial Intelligence', signed by an impressive number of scientific,

industrial, and intellectual personalities including (among others) Elon Musk and Stephen Hawking, was one of the first to highlight multiple ethical challenges in both the long and the short runs (Russel, Dewey, and Tegmark, 2015). While prominently featuring the spectre of a super-intelligence allegedly bound to replace humans, the letter also pointed to more concrete ethical concerns relative, for example, to autonomous weapons and surveillance cameras. Other moral dilemmas included new versions of the classic 'trolley problem', adapted to driverless cars choosing the lesser of two evils, between killing pedestrians or passengers in case of an accident. In the letter, risks were addressed alongside grossly inflated opportunities – like 'eradication of disease and poverty'.

Although the document seemed to situate such issues and thought experiments in the future, its authors fired the opening salvo of a series of ethical charters, codes of conduct, and guidelines that were to emerge in the following years. Private companies, public institutions, and non-governmental organisations tried to systematise and prioritise values and notions related to ethical AI. To get a sense of the diversity of institutional actors that have addressed the issue, one needs only note that in 2020 a pledge for ethical AI was signed between IBM, Microsoft, and (more surprisingly) the Vatican (Copestake, 2020). But how many charters are there, who are their authors, and what are their key contents? A study published in 2019 identified 84 charters, mostly written since 2016 (Jobin, Ienca and Vayena, 2019); the crowdsourced AI Ethics Guidelines Global Inventory, maintained by the association Algorithm Watch (2020), counted 173 in April 2020, most of them developed between 2018 and 2019. Hagendorff (2021, p.3) notices that newer guidelines were inspired by older ones, so that especially the most recent codes are often echoes of each other. This justifies taking the research of Jobin, Ienca and Vayena (2019) as representative of the ethical AI landscape.

Although no single principle is common to all the documents in their corpus, the authors identify five most recurrent themes: *transparency, justice and fairness, non-maleficence, responsibility,* and *privacy*. These ethical domains are understood differently in each of the documents. Some guidelines interpret *transparency* as a technical standard and recommend the adoption of open-source code, while others see it as a set of best practices to avoid decision-making by opaque automatic systems, or even as an economic issue concerning the declaration of funding bodies for automated systems. Similarly, *justice* is a multifaceted notion that may include the need to limit bias so that algorithms do not systematically discriminate against certain individuals or groups; the possibility of appealing against a decision taken by an algorithm; or the possible impact of automated systems on labour markets. Regarding this last point, most of the existing ethical guidelines limit themselves to highlighting the dangers of technological labour displacement – a long-established controversy that dates back to the first industrial revolution. In turn, the principle of *non-maleficence* (curiously much more common than 'beneficence') often coincides with the classic liberal 'no-harm' principle and states the need to implement AI models that are incapable of causing inconvenience to human beings. Moreover, the core value of *responsibility* is paramount in several AI ethical

guidelines, especially important in relation to legal accountability and liability in case of accidents involving autonomous systems. Asking 'who is responsible' implies envisioning AI not as a technological artefact replacing human decision-making, but rather as a tool distributing responsibility for the decision between human and non-human actors. Finally, the value of *privacy* protection is characterised by popular portmanteau notions such as 'privacy by design', although promising new technical solutions are emerging, notably 'differential privacy' which allows describing patterns in a dataset while withholding information about individuals in it.

According to Hagendorff (2020, p.103), the principles of transparency, responsibility, privacy, and to some extent fairness can be 'most easily operationalised mathematically and thus tend to be implemented in terms of technical solutions'. Almost all extant guidelines envision ethical issues as purely technological problems that machine learning itself can resolve. In part, this is the result of educational approaches that traditionally devalue any non-technical inputs. Even today's expanding attempts to include AI ethics in the training of computer scientists and engineers follow an approach that Raji, Scheuerman, and Amironesei (2021) call 'exclusionary pedagogy'. In this teaching methodology, ethics is distilled for computational uses, but there is no deeper epistemological engagement with other ways of knowing that would benefit ethical thinking, most notably those based on social sciences. Such limited and uni-vocal computational approach to dilemmas stemming from the collective adoption of AI systems results in an unhealthy 'promotion of the ideal of *ethical unicorns* or *tech saviours*' (Raji, Scheuerman, and Amironesei 2021, p.516).

Reducing ethics principles to their technical operationalisation fails to take into account the broader socio-economic context, the geographical environment, the institutional settings and the networks of relationships in which AI systems are embedded, and how they may differentially affect individuals, groups and organisations. This limitation has momentous implications. For example, it has been claimed that booming research on 'algorithmic fairness' that aims to devise technical solutions to bias and discrimination, misses the socially constructed nature of protected attributes such as gender and race, taking them as fixed categories, while they are instead evolving institutional and relational phenomena embedded in historically specific settings (Denton et al., 2020; Hanna et al., 2020). It is therefore important to adopt a more holistic view, looking at the whole systems of which AI (and the ethics guidelines developed for it) is part. In this respect, there is scope for significant input from digital social sciences. While AI technologies play an increasing role in shaping our individual lives and society, they are themselves the product of social, economic, political, and historical factors. In what follows, we outline how some of the methods and tools that the social sciences have developed in the last two decades to recognise and analyse the human contexts of data and technologies can provide novel and fruitful insights to fill the gaps in current approaches to AI ethics.

WHOSE VALUES ARE PRIORITISED IN AI ETHICS GUIDELINES?

When extending our gaze to encompass these broader contexts, drawing on insights from the social sciences, it appears that the geographical distribution of the issuers of ethical guidelines studied by Jobin, Ienca and Vayena (2019) displays hotspots and hubs in Europe, the United States, and to a lesser extent Japan and India, all countries that are investing heavily in AI development (although a limitation of the study is lack of coverage of China), while the Global South is virtually absent from these contributions. Put differently, the geographical distribution of AI ethics charts overlaps with the geographical distribution of the marketing of AI solutions more generally.

Such an overlap involves the risk of potential conflicts of interest. Why sustain principles that, if effectively applied, may interfere with the ongoing exponential path of technological development and commercialisation? For example, as far as privacy protection is concerned, full-fledged efforts to minimise risks may amount to limiting personal data collection and monetisation, thereby severely hindering data access and therefore dampening the potential of machine learning. The temptation is to focus exclusively on ethical principles that do not fundamentally conflict with free market ideologies and companies' profit objectives.

What emerges here is the permeability of the field to industrial interests, little addressed in extant ethics guidelines. In this respect, AI ethics follow suit to a grand tradition in engineering and computing research, as well as in education, aiming to address industry needs (Tomayko, 1998). Many of the existing guidelines have been designed by, or with the contribution of, technology companies pursuing self-regulation. According to Whittaker (2021, p.54), corporate actors recur to ethical AI to 'co-opt and neutralise critique':

> They do this in part by funding and elevating their weakest critics, often institutions and coalitions that focus on so-called AI ethics, and frame issues of tech power and dominance as abstract governance questions that take the tech industry's current form as a given and AI's proliferation as inevitable. […] Such approaches make great PR. They also serve to cast elite engineers as the arbiters of 'bias,' while structurally excluding scholars and advocates who don't have computer science training…

The implicit or explicit role of technology companies in shaping AI ethics appears to fill in (while actually contributing to) the current regulatory vacuum. 'Ethics washing' defuses attempts to develop and deploy potentially much more restrictive regulation, involving legally mandated standards and enforcement mechanisms. The underlying assumption is that 'innovation is an unvarnished and unmitigated good' that government regulation would stifle to the detriment of all, and that 'the tech industry can formulate appropriate ethical norms for AI and can be trusted to ensure that AI systems will duly adhere to those standards' (Yeung, Howes and Pogrebna, 2020, p.78). That the latter assumption does not hold is proven by the almost complete absence of discussions about

forms of effective enforcement of ethical standards in these charts and, as Hagendorff (2020, p.99) puts it, by the lack of any significant actual impact on machine learning and AI. Even when ethical principles are operationalised via specific tools, they fall short to address existing imbalances and limitations. A recent search of 169 ethics documents found only 39 included AI ethics tools to put principles in practice, for example lists of best practices, checklists, and adapted software applications. This study also identifies two major gaps in the actual use of these tools, namely exclusion of key stakeholders and lack of external auditing (Ayling and Chapman, 2022).

ETHICS FOR WHAT? DEPLOYMENT VS PRODUCTION OF AI

Having recognised the embeddedness of AI ethics discourses in the commercial interests of technology companies, we must also acknowledge that AI is industrially *produced* – that is, it results from the mobilisation of human and natural resources, requiring appropriate corporate structures, capital investments and adequate institutional systems enabling its development. What is, then, the human, economic and social context within which AI is produced? Who contributes to it, and for what rewards? What are its production costs – seen from the viewpoint of society as a whole rather than the individual producer? Of note, virtually none of these questions are addressed in AI ethics charts, which implicitly assume that voice assistants, recommendation engines, and self-driving vehicles raise ethical concerns mainly when consumers use them. Like other areas of research on the socio-political impacts of AI, the focus of ethics has been largely on the phase of its *deployment* on markets, on the possible effects of its being put to use widely in society. This has triggered highly disembodied debates on applications that are only being theorised or at most pilot-tested, but are not currently operating, while leaving aside the – already present and visible – socio-political and economic context in which AI is being developed now.

The example of autonomous vehicles shows the importance of the production processes behind AI, well ahead of its subsequent deployment (Tubaro and Casilli, 2019). In addition to engineers, software developers and designers, there is a need for ‘vehicle operators’ or ‘safety drivers’ who travel inside the car, monitor the trip, provide feedback to the technical teams, and are expected to take control of the car if necessary. But if vehicle operators are visible to all – to the displeasure of manufacturers – an army of other, more hidden workers are required to support the development of so-called self-driving cars. An autonomous vehicle is endowed with computer-vision algorithms that must be able to recognise what is going on – for example, if a pedestrian is crossing the street. But how does the machine figure out what a pedestrian looks like? As in all machine learning, it will need large sets of data as examples, and the images routinely taken by radars, cameras, and sensor devices mounted on robot cars, provide precisely this. To be useful, these images first need to be labelled: in a road traffic photo, the computer needs indications of what elements are pedestrians, and what are, instead,

bikes, traffic lights, or buses. It is here that human workers intervene – people who are paid to identify and tag everything that can be seen in each image. The job is necessarily huge because, as discussed, the machine can only learn from large amounts of data, that would be tedious and lengthy to annotate if just one or few workers were in charge. The solution consists in fragmenting these large batches into many short, one-shot tasks, and allocate them to many human providers, each of whom will do just one or few. This part of human work, indispensable for developing autonomous cars, is performed off the street by myriad so-called 'micro-workers' who operate remotely. Thus, one might ask under what conditions these human workers are recruited, whether their remunerations are fair, how they are managed, what rights they enjoy. In addition to this labour force, self-driving cars consume natural resources, including at the very least the minerals and metals necessary to build them, and the energy they consume – both to actually circulate and before that, to run heavy computations on powerful servers. In this sense, they raise questions in terms of their sustainability – an issue that occurs rarely in charts and guidelines and that has relatively little surfaced in AI ethics debates so far (Müller, 2021).

In terms of production-related ethical issues, autonomous cars are not an isolated case, but rather exemplify trends that are observed throughout the AI industry. We can generalise from this example and provide more background on the two main issues that arise regarding, respectively, human labour and the environment.

AI PRODUCTION: MICRO-WORK AS AN ETHICAL PROBLEM

If the AI industry values highly – and compensates accordingly – its visible workforce of software developers, data scientists, and computer engineers, it tends to be silent on the lower-level micro-workers who perform the indispensable, yet less qualified, often repetitive, and unchallenging support tasks that are needed to keep the system going. These invisibilised humans intervene at different stages of the production process of a machine learning model: its initial training (data generation and annotation), verification of its outputs once the model has been deployed and put to use, and sometimes real-time correction or 'impersonation', whereby humans step in to perform tasks that malfunctioning AI systems fail to complete (Tubaro, Casilli and Coville, 2020).

Amazon's micro-work specialised platform, Mechanical Turk, was the first to popularise human-powered micro-tasking for AI in the mid-2000s. It takes its name from a famous eighteenth-century chess-playing automaton in Ottoman disguise. This AI ahead of its time could supposedly reproduce the cognitive processes of a human chess player, but it was in fact a hoax, manoeuvred by a human operator hidden inside its gears. Today, it serves as a metaphor of the human-in-the-loop principle that governs AI production. The involvement of human workers to prepare, test, and sometimes pose as autonomous systems, still follows the same logic but on a much larger scale. Not one, but hundreds of thousands of freelance workers perform so-called human intelligence tasks (HITs) which are mostly straightforward or even trivial for humans, yet hard to automate.

Unironically, Amazon described the result of this mass micro-work as ‘artificial artificial intelligence’.

Since the launch of Mechanical Turk, many more actors have appeared on this market. Several international platforms sell micro-working services on demand to clients worldwide, such as the Australian Appen, the Americans Oneforma and Remotasks, or the German Clickworker. The largest technology multi-nationals have created their own micro-working services on which they act as monopsonists, such as Microsoft with UHRS (Universal Human Relevance System) and Google with RaterHub. Micro-working intermediaries also include companies that operate as BPO (Business Process Outsourcing) vendors, hiring workers in countries where they are cheaper, for example the Indian Playment and the American Cloudfactory, which recruits workers in Kenya and Nepal. Some of these companies and platforms are very large, others smaller and sometimes specialised, for example the French IsAHit and Wirk. Some specialised start-ups have been bought by larger actors, for example the Californian Mighty AI, dedicated specifically to the needs of the car industry and incorporated in 2019 by Uber Technologies.

AI developers and technology companies can use these and similar intermediaries to recruit disposable workers for their projects. In addition to image annotation, other commonly found tasks involve taking selfies (to feed face recognition algorithms), recording one’s voice (to provide soundbites to train voice assistants), transcribing from sound files, translating and editing short bits of text (for language-processing applications of AI). This work is very important for a multitude of technology uses. The producers of search engines such as Bing and Google use micro-workers’ services to check the relevance of search engine results (is ‘Dijon mustard’ the same as ‘mustard of Dijon’?), while social media and other websites often need micro-workers for content moderation (flagging offensive, illegal or otherwise unwelcome texts, images, or videos).

The size of the phenomenon has increased over time. According to Tubaro, Le Ludec and Casilli (2020), there are about 260,000 micro-workers in France only, who use a number of national and international platforms. A more recent study includes micro-workers among the over 163 million online workers estimated worldwide (Kässi, Lehdonvirta and Stephany, 2021).

The conditions under which this type of work is performed are problematic. Contingent micro-workers are rarely formally employed and do not figure on the company's payroll. In most cases, they are bound to platforms only via membership or participation contracts that resemble general Terms of Use, no different from those routinely agreed to by consumers of manifold internet services or mobile applications. For this reason, they are exposed to the volatility of the market and receive no form of social protection. Further, they do remote work that most often can be performed from anywhere. The result is openness to worldwide competition and over-supply of labour, which drives down remunerations. While some platforms (such as Clickworker)

recommend paying at minimum wage, the common practice of paying by piecework rather than by the hour, makes any controls difficult. In practice in the worst cases, a micro-task can be paid as little as a few cents. In their multi-country, multi-platform study, Berg et al. (2018) find that on average, micro-workers earn only US$ 3.31 per hour (US$ 4.43 if the time to search for tasks is not counted), which is well below the minimum wage of most countries.

Despite the low earnings that micro-tasks offer, workers are often motivated by economic necessity. Among French microworkers, Casilli et al. (2019) note that people with low income, and those living below the poverty line, are over-represented. Women with young children who have a main part-time employment and use micro-tasks to supplement an otherwise meagre income, are commonly found in a country like France (Tubaro et al., 2022). To date, there is no clear evidence that these workers can derive additional advantages from this activity – such as additional skills that may position them favourably on the labour market in the future.

All this counters some the basic principles outlined in most of the above-discussed AI ethics charts, notably fairness and justice (to the extent that AI producers materially benefit at the expense of micro-workers), transparency (to the extent that the human contribution to AI is invisibilised) and even privacy (as workers are sometimes asked to provide personal data, such as selfies and voice recordings, for the needs of dataset creation). There is a deep contradiction here insofar as the industry cannot meet its own standards – however light and vaguely defined they may be, compared to potentially stricter legal regulation. The main problem is that those standards were set with only the deployment of AI in mind, and without any consideration for production.

GEOGRAPHY OF DATA LABOUR

If the very existence of invisibilised, poorly paid, and unglamorous micro-work is in itself problematic, its geographical distribution further challenges today's approaches to AI ethics. As discussed above, guidelines tend to be published in high-income countries, whilst the geography of data labour such as annotation and filtering of data necessary to produce AI can be construed as its flipped image. Although the micro-workers of Global North countries such as the United States (Difallah, Filatova and Ipeirotis, 2018) and France (Casilli et al., 2019) have attracted most research attention so far, the majority are located in the Global South. Extant evidence suggests that their global distribution reproduces legacy inequalities at the planetary level in terms of wealth, power, and geographic influence (ILO, 2021, p. 45). These results are mostly based on data from English-speaking platforms and highlight clear linkages with former British colonies like India (see also Gray and Suri, 2019) or former zones of U.S. hegemony, such as the Philippines (see also Roberts, 2019). Inspired by these approaches, new research is currently being carried out on the Francophone world, which extends from the flourishing French AI industry to (mostly French speaking) African countries

like Madagascar, Cameroon, Mali, Senegal, Morocco, but also Egypt. Another large, and yet largely unexplored, reservoir of micro-workforce is Latin America, with countries like Venezuela, Argentina (Miceli, Posada and Yang, 2022), and Brazil (Grohman and Fernandes Araújo, 2021)playing a very active role in this international market.

Fig. 1. Global flows of annotated data from micro-work providers to AI solution providers. Source: authors' elaboration.

Figure 1 schematically represents the global flows through which, based on the evidence presently available, micro-work feeds the development of AI. Latin American workers serve technology producers in North America (especially in the United States), and to a lesser extent Russia and Europe. From South and South-East Asia, work is directed primarily toward North America, but also, in part, toward China. Africa serves both Europe and China. Internal flows are important within China, though little is known about them, and to a lesser extent, within Europe (with some East-West flows) and within the United States. While the map should only be taken as indicative and qualitative, as the exact size of the flows still needs to be documented in many cases, it is sufficient to highlight an additional major gap that current approaches to AI ethics fail to fill: the severe under-representation of the Global South in the creation of charts and guidelines, combined with its over-representation in the 'human' supply chain that produces AI,

undermines pluralism and cultural awareness, while also further contributing to limiting workers' voice and invisibilising their presence.

WORKING ON THE MINERALOGICAL LAYER

If we want to produce an AI that respects the human environments where labour is produced, we have to conceive it as a technological system that also respects the natural environment where resources are put to use. Although few extant ethical guidelines explicitly mention sustainability as a principle (see above), attention to the environmental costs of AI has been growing in recent years. This challenge has been addressed in two fundamentally different ways: by assessing the carbon footprint of computational processes and by denouncing the extractivistic logic on which the tech industry is predicated.

In the last few years, efforts have been made to measure the environmental cost of machine learning. In a path-breaking paper, Strubell, Ganesh and McCallum (2019) estimated the carbon dioxide emissions to train a single large Natural Language Processing (NLP) transformer model at 626,155 lbs. This is nearly five times the combined emissions resulting from the manufacture and lifetime of a car (including its fuel), or more than 50 times the emissions of one human being per year.

Increasingly, AI researchers assess the energy use of their tools via different trackers and impact measurements. The ultimate goal is to suggest actionable recommendations to reduce carbon dioxide emissions (Bannour et al., 2021). Major companies, such as Alphabet, have developed more efficient ways to cool their data centres, and are investing heavily in green energy sources (Evans and Gao, 2016). Other solutions focus on developing 'green algorithms', efficient architectural settings, and smaller-sized models to train AI (Cai et al., 2020). According to a 2020 report issued by the consulting firm Capgemini 'AI-enabled use cases have helped organisations reduce GHG emissions by 13% and improve power efficiency by 11% in the last two years. AI use cases have also helped reduce waste and deadweight assets by improving their utilisation by 12%' (Capgemini Research Institute, 2020, p.2).

While applauding these commendable efforts to measure, highlight and reduce environmental impacts, it must be recognized that once again, they rely on a self-regulating approach and assume that tech industries are able and willing to stick to it. Along these lines, one would have to assume that the very technological systems that pollute and waste natural resources also entail the potential to mitigate their own adverse effects. Yet, over-reliance on abstract categories such as 'environment', 'climate', or 'energy' in this body of literature obfuscates the concrete human and material substratum that supports the transformation of natural resources into data-intensive AI solutions.

To go deeper into the analysis of the materiality of AI, its reliance on natural resources, and its linkages to contemporary capitalism, other researchers have developed

a more radical theoretical perspective centred on the notion of digital extractivism. Broadening the original meaning of extraction as the plundering of natural resources, they use the notion to describe how the operations of capital interact with and draw on multiple sectors of human, political, economic and social activity. This extractivist logic is omnipresent in contemporary capitalism, spanning not only traditional sectors like logistics and agriculture, but also immaterial activities like finance and, of note, data production on platforms (Mezzadra and Neilson, 2019). In this respect, AI is a new front of extraction because algorithms and data would not function without key minerals and metals that serve to build their core hardware components. To produce batteries for mobile devices and autonomous vehicles, minerals such as cobalt, nickel and lithium, are marshalled in Nevada and most prominently in countries like Madagascar, Bolivia, Argentina and Chili. Likewise, the high-income countries that produce semiconductors, such as Taiwan and South Korea, procure tin from Indonesia (Unknown Fields, 2016; Crawford, 2021, p.30).

This constitutes what Crawford (2021, p.32) calls ‘the mineralogical layer’ of AI – the literal bedrock of the informational infrastructure that nourishes intelligent solutions – to which a ‘logistical layer’ must also be added to account for the environmental costs of moving minerals, metals, fuel, hardware, and final products across countries.

The extraction of raw materials is known to prompt conflicts, due to displacement of the local populations, disruptions to their traditional activities and ways of life, and labour struggles among miners. The case of the Democratic Republic of Congo is a clear example of how political instability may mirror the violence of coltan mining. Conflicts can escalate at international level: the decade-long US-China trade conflict has revolved around economic protectionism in favour of local technology and telecommunications companies as well as embargo measures on rare earths, of which both countries detain the largest reserves in the world (Vercellone, 2020).

Figure 2 shows how a map of the locations where most natural resources are extracted for the AI industry overlaps significantly with the trajectories of annotated data flows represented in Figure 1. Together, the two figures highlight how several groups of workers from the Global South are mobilised for the AI industry of the Global North, comprising both the data micro-workers discussed earlier, and all those who work in mining, transport, and electronic waste management. In the words of Fuchs (2016, p.17) ‘just like the labour of workers in the periphery during earlier stages of imperialism, the production of information and information technology is part of an international division of labour’. The cycle of production of the information and contents circulating from remote servers to our screens does not only revolve around activities that produce and transform data, but also around the labour of extraction and transportation of the minerals used in electronics, semiconductors, and batteries.

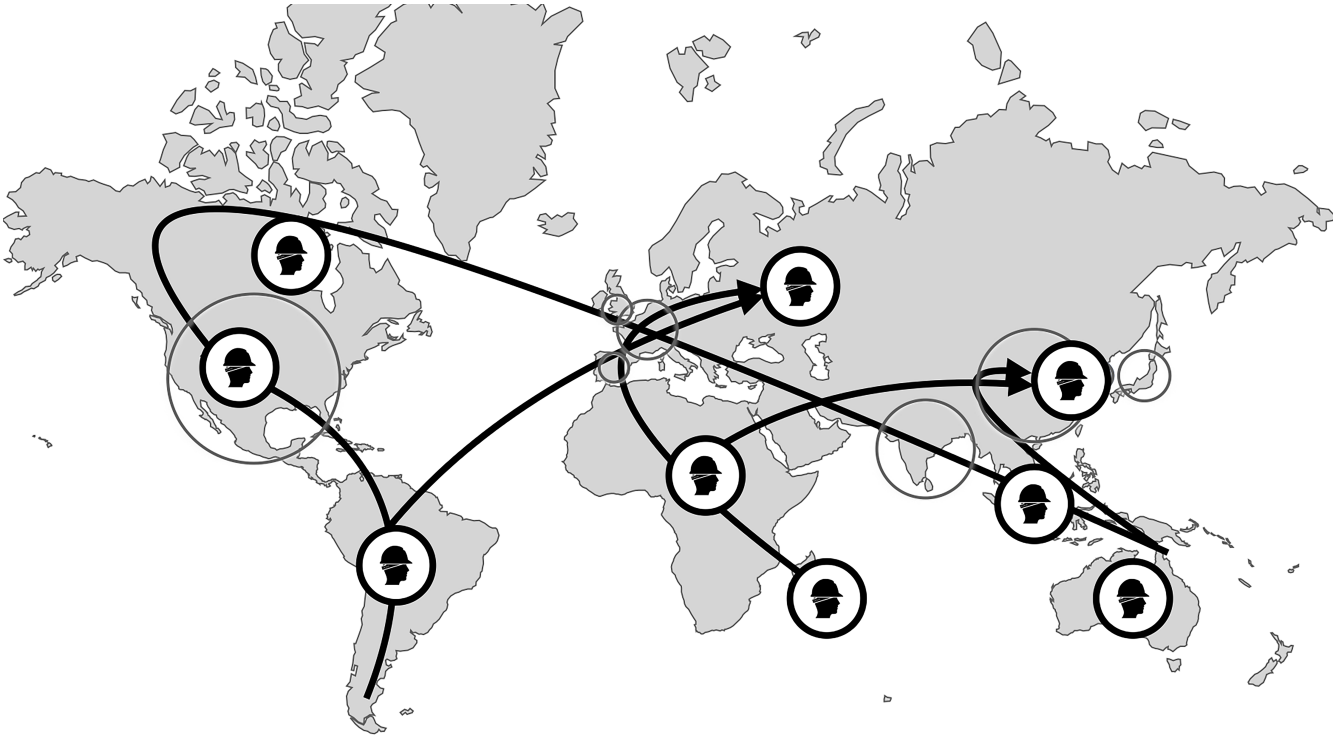

Fig. 2. Hotspots of raw materials' extraction intersect flows of annotated data from micro-work providers to AI solution producers. Source: authors' elaboration.

'END-TO-END' ETHICAL AI: FROM DEPLOYMENT TO PRODUCTION

By looking at the social and economic dimensions of today's AI production processes, we have highlighted major limitations in existing technology-oriented approaches that would escape our gaze if we confined our attention to the principles highlighted in current public debates and guidelines. But this perspective also suggests how to overcome the limitations of current AI ethics. In accordance with the approaches developed among others by Crawford (2021) and Kazimzade and Miceli (2020), an emerging alternative research programme looks at both ends of AI (Casilli, 2020), its consumption by final users and its production by annotators and other workers. This approach allows construing AI differently, taking into account the full social, economic, political, and not only techno-scientific context in which it emerges. Beyond the now popular representation of AI as a data-intensive technique, it raises the essential question of the conditions under which such data is produced.

If we want to propose an *end-to-end ethical AI*, we must take into account the conditions of production of the data, tools, and equipment used to create and to market its artefacts. In this respect, we should apply the same kind of ethical reasoning that is already in use with a number of other consumer products. For example, producers of footwear or of processed food are not automatically thought to trade ethically if they avoid discriminating their consumers by gender, geographical origin, or some other

attributes. They are expected to respect the rights of workers and to offer them decent working conditions, and/or to minimise the environmental impact of their production. Likewise, an ethical AI cannot only be unbiased in its deployment but must also alleviate the negative externalities of its production processes—both in human communities and within natural environments.

CONCLUSION

Despite largely shared emphasis on the values of transparency, confidentiality, justice, accountability, and non-maleficence, the ethics of AI is still characterised by divergent and conflicting trends that mirror underlying economic and political interests. Primarily promoted by technology companies and other institutions and organisations located in AI-producing countries, guidelines and similarly mild self-regulatory tools constitute a way for corporate actors to avoid potentially more constraining governmental regulations. Further, today's approaches to AI ethics focus only on the deployment of automation, largely ignoring the material infrastructure and environmental costs of its production, together with the fragmented and underpaid labour of millions of data workers needed to make it possible.

We have shown that a social science-informed perspective that fully takes into account the human, social, political and economic contexts of today's AI technologies can provide novel insights and suggest directions for future action. What if the rights of remote workers operating both on the mineralogical and on the informational layers of AI production were protected, if they had the possibility to resist unsuitable conditions, if they had voice to protest against, or to refuse to contribute to, any technological system that they would consider ethically problematic? In such a scenario, the focus of AI ethics would shift from one that constantly coaxes the interests of producers-owners to one that advocates the moral agency of producers-workers. A stronger, more fundamental approach to AI ethics – or as Casilli (2020) calls it, a 'truly ethical' AI – requires recognising the currently invisible work of preparation, verification, and impersonation of automated systems, and should consequently provide workers with methods to protect themselves. The same principles could be extended to all other workers directly and indirectly involved in the supply chain of today's computing devices.

REFERENCES

ALGORITHM WATCH (2020). *AI Ethics Guidelines Global Inventory* [online]. Available from: https://inventory.algorithmwatch.org/ [Accessed 05/02/2022].

AYLING, J. and CHAPMAN, A. (2022). Putting AI ethics to work: are the tools fit for purpose? *AI Ethics*, 2, pp. 405–429. Available from: https://doi.org/10.1007/s43681-021-00084-x

BANNOUR, N., GHANNAY, S., NÉVÉOL, A. and LIGOZAT, A.-L. (2021). Evaluating the carbon foot- print of NLP methods: a survey and analysis of existing tools. EMNLP, Workshop SustaiNLP, Punta Cana, Dominican Republic. Available from: https://doi.org/10.18653/v1/2021.sustainlp-1.2

BERG, J., FURRER, M., HARMON, E., RANI, U. and SILBERMAN, M.S. (2018). *Digital labour platforms and the future of work: Towards decent work in the online world.* Report. Geneva: International Labor Organization (ILO).

BUOLAMWINI, J. and GEBRU, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* vol. 81 of *Proceedings of Machine Learning Research*, pp. 77-91.

CAI, H., GAN, C., WANG, T., ZHANG, Z. and HAN, S. (2020). Once-for-all: Train one network and specialize it for efficient deployment. *International Conference on Learning Representations (ICLR)*.

CAPGEMINI RESEARCH INSTITUTE (2020). *Climate AI. How Artificial Intelligence can Power Your Climate Action Strategy.* Available from: https://www.capgemini.com/research/climate-ai/ [Accessed 21/02/2022].

CASILLI, A.A. (2021) Qu'est-ce qu'une intelligence artificielle "réellement éthique"? In *Proceedings of the TESaCo Conference*. Paris: Institut de France, pp. 71-78.

CASILLI, A.A., TUBARO, P., LE LUDEC, C., COVILLE, M., BESENVAL, M., MOUHTARE, T. and WAHAL, E. (2019). *Le Micro-Travail en France. Derrière l'automatisation, de nouvelles précarités au travail?* Report. Paris, Digital Platform Labor (DiPLab) project.

CASILLI, A.A. (2015). Four theses on digital mass surveillance and the negotiation of privacy. *8th Annual Privacy Law Scholar Congress*, June 4, Berkeley, CA, USA. Available from: https://halshs.archives-ouvertes.fr/halshs-01147832 [Accessed 21/02/2022].

COPESTAKE, J. (2020). AI ethics backed by Pope and tech giants in new plan. BBC News. Available from: https://www.bbc.com/news/technology-51673296 [Accessed 21/02/2022].

CRAWFORD, K. (2021). *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*. London and New Haven: Yale University Press.

DENTON, E., HANNA, A., AMIRONESEI, R., SMART, A., NICOLE, H. and SCHEUERMAN, M.K. (2020). Bringing the people back in: contesting benchmark machine learning datasets. *Proceedings of ICML Workshop on Participatory Approaches to Machine Learning.*

DIFALLAH, D., FILATOVA, E. and IPEIROTIS, P. (2018). Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. New York: ACM, pp. 135-143.

EVANS, R. and GAO, J. (2016). DeepMind AI reduces energy used for cooling Google data centers by 40%. *The Keyword Blog*. Available from: https://blog.google/outreach-initiatives/environment/deepmind-ai-reduces-energy-used-for/ [Accessed 18/02/2022].

FJELLAND, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, 7(10), pp. 1-9.

FUCHS, C. (2016). Digital labor and imperialism. *Monthly Review*, 67(8), pp. 14-24.

GRAY, M.L. and SURI, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Boston and New York: Houghton Mifflin Harcourt.

GROHMANN, R. and FERNANDES ARAÚJO, W. (2021). Beyond Mechanical Turk: The work of Brazilians on global AI platforms. In: VERDEGEM, P. (ed.) *AI for Everyone? Critical Perspectives*. London: University of Westminster Press, pp. 247-266.

HAGENDORFF, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines,* 30, pp. 99-120.

HAGENDORFF, T. (2021). Blind spots in AI ethics. *AI Ethics* [online]. Available from: https://doi.org/10.1007/s43681-021-00122-8

HANNA, A., DENTON, E., SMART, A. and SMITH-LOUD, J. (2020). Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York: ACM, pp. 501-512.

INTERNATIONAL LABOR ORGANIZATION (ILO). (2021). *The Role of Digital Labor Platforms in Transforming the World of Work*. Flagship Report. Geneva: ILO. Available from: https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/---publ/documents/publication/wcms_771749.pdf

JOBIN, A., IENCA, M. and VAYENA, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, pp. 389-399.

KÄSSI, O, LEHDONVIRTA, V. and STEPHANY, F. (2021). How many online workers are there in the world? A data-driven assessment [version 4; peer review: 4 approved]. *Open Research Europe* [online], 1(53). Available from: https://doi.org/10.12688/openreseurope.13639.4

KAZIMZADE, G. and MICELI, M. (2020). Biased priorities, biased outcomes: Three recommendations for ethics-oriented data annotation practices. In *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society Proceedings*, New York: ACM, p. 71.

MEZZADRA, S. and NEILSON, B. (2019). *The Politics of Operations. Excavating Contemporary Capitalism*. Durham: Duke University Press.

MICELI, M., POSADA, J. and YANG, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? In *Proceedings of the ACM Human-Computer Interaction,* New York: ACM, 6(34), pp. 1-14. Available from: https://doi.org/10.1145/3492853

MÜLLER, V.C. (2021). Ethics of artificial intelligence and robotics. In ZALTA, E.N. (ed.), *The Stanford Encyclopedia of Philosophy* [online]. Available from: https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/

RAJI, I. D., SCHEUERMAN, M. K. and AMIRONESEI, R. (2021). “You can’t sit with us”: Exclusionary pedagogy in AI ethics education. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: ACM, pp. 515-525.

ROBERTS, S.T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media.* London and New Haven: Yale University Press.

RUSSELL, S., DEWEY, D. and TEGMARK, T. (2015). Research priorities for robust and beneficial Artificial Intelligence. *AI Magazine*, 36(4), pp. 105-114.

STRUBELL, E., GANESH, A. and McCALLUM, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL),* Florence, IT.

TOMAYKO, J.E. (1998). Forging a discipline: An outline history of software engineering education. *Annals of Software Engineering,* 6(1-4), pp. 3-18.

TUBARO, P., COVILLE, M., LE LUDEC, C. and CASILLI, A.A. (2022). Hidden inequalities: the gendered labour of women on micro-tasking platforms. *Internet Policy Review* [online], 11(1). Available from: https://doi.org/10.14763/2022.1.1623

TUBARO, P., CASILLI, A.A. and COVILLE, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data and Society* [online], 7(1). Available from: https://doi.org/10.1177/2053951720919776

TUBARO, P., LE LUDEC, C. and CASILLI, A.A. (2020). Counting ‘micro-workers’: societal and methodological challenges around new forms of labour. *Work Organisation, Labour and Globalisation*, 14(1), pp. 67-82.

TUBARO, P. and CASILLI, A.A. (2019). Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics*, 46(3), pp. 333-345.

TUBARO, P., CASILLI, A.A. and SARABI, Y. (2014). *Against the Hypothesis of the End of Privacy: An Agent-Based Modelling Approach to Social Media.* New York: Springer.

UNKNOWN FIELDS. (2016). *The Breastmilk of the Volcano. Bolivia and the Atacama Desert Expedition*. London, UK: Actar Publishers.

VERCELLONE, C. (2020). Big-data e Free Digital Labor nel capitalismo delle piattaforme: un nuovo estrattivismo? *Proceedings of the conference ‘L’enigma del valore. Il digital labor e la rivoluzione tecnologica’*. Milan: Effimera, pp. 9-24.

WACHTER, S., MITTELSTADT, B. and FLORIDI, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), pp. 76-99.

WHITTAKER, M. (2021). The steep cost of capture. *Interactions*, 28(6), pp. 50-55.

YEUNG, K., HOWES, A. and POGREBNA, G. (2020). AI governance by human rights–centered design, deliberation, and oversight: An end to ethics washing. In DUBBER, M.D., PASQUALE, F. and DAS, S. (eds.), *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press, pp. 77-106.