AI AND ETHICS: MOVING TOWARDS A FUTURE WHERE AI BENEFITS – RATHER THAN DESTROYS – HUMANITY

Adam James Fenton[1], Dagmar Monett[2]
Coventry University UK (Centre for Trust Peace and Social Relations)[1]
Berlin School of Economics and Law (Computer Science Dept.)[2]

ABSTRACT

2023 may be remembered as the year that Artificial Intelligence (AI) hit the headlines for its potential to transform economy, society, and humanity. With the launch of OpenAI’s ChatGPT, the wider public experienced first-hand what is it like to interact with an AI that shows remarkably human-like responses in one-to-one conversations. This is clearly a technology with the potential for profound (both negative and positive) impacts on all aspects of human activity from business to politics, justice, education, arts, science, medicine, and virtually any other field of human endeavour. The challenges, as many academics, practitioners, observers, and some legislatures have realised, are in the capabilities of the underlying technology and in regulating AI and ensuring that its deployment and use are ethical, responsible, and trustworthy; that it is used for good rather than evil, to benefit humanity rather than destroy or damage it. However, given both the complexity of the technology itself and the competing vested interests of the many interested parties involved, this is a task which is much easier said than done. Indeed, some commentators have noted that the cursory nature with which ethics has been approached by some actors in the AI regulation debate has been nothing more than “ethics washing” (Metzinger, 2019a, 2019b; van Maanen, 2022) or even “ethics theater” (Cath & Jansen, 2021). Simply articulating what is meant by “ethical AI” is a task freighted with intercultural challenges, as one may ask “whose ethics are we referring to?” Ethical systems vary through history and across cultures, making it difficult, if not impossible, to articulate a “universal” form of AI ethics. This paper examines the meta-ethical challenges of AI and its regulation, and offers novel recommendations to improve discussions and outcomes around the ethics and regulation of contemporary AI.

Keywords: Artificial intelligence (AI); AI ethics; AI regulation; intercultural communication and AI; explainable AI; technology; cybersecurity.

INTRODUCTION

Beginning in late 2022, the public launch of chatbots like OpenAI’s ChatGPT, Google’s Bard and Anthropic’s Claude, has sparked enormous interest and renewed debate around both Artificial Intelligence (AI), in general, and intelligent capabilities in machines (or algorithms, robots, etc.), in particular. This is manifested in an explosion in media and scientific publications[2], as well as public discussion, around the capabilities, and regulation, of AI; some of it measured and sound, some of it wildly speculative and alarmist. To put this discussion in context, two points can be noted at the outset. First, AI is not a

[1] This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101029232

[2] One study found: “From 2010 to 2021, the total number of AI publications more than doubled, growing from 200,000 in 2010 to almost 500,000 in 2021.” (Maslej et al., 2023, p. 24).

new field of computer science that has suddenly burst onto the scene, it “has been the subject of continuous research for more than half a century” (Wooldridge, 2020, p. 5). Second, throughout those decades it has gone through several boom and bust cycles (also known as AI Springs and AI Winters), where “researchers have repeatedly claimed to have made breakthroughs that bring the dream of intelligent machines within reach, only to have their claims exposed as hopelessly overoptimistic in every case” (Wooldridge, 2020, p. 5).

Whether the current interest in AI is destined for yet another “bust” is yet to be seen. What is perhaps different to previous cycles is that the wider public has now directly experienced interacting with large language models (LLMs) indirectly, like the ones used in generative AI-based applications like current chatbots, capable of producing language, images and audio that is, in some cases at least, remarkably human-like and arguably capable of passing Turing’s famous “imitation game” test[3] (Turing, 1950). While these LLMs appear to be capable of holding “intelligent” conversations across virtually any topic, it must be underlined that they are not infallible and have been repeatedly shown to be prone to factual errors (Alkaissi & McFarlane, 2023).

Much of the fear and fascination around AI stems from images and stereotypes, lodged in the collective cultural psyche, of intelligent machines that rise up and destroy humanity. This is based on a fiercely contested concept of the ‘singularity,’ a theoretical point in the future where a recursive ‘intelligence explosion’ occurs, creating AI that is so advanced it will be able to set its own goals, and may be beyond human control (Chalmers, 2010; Good, 1966). Others point out that “most AI researchers … are very sceptical about the Singularity, at least for the foreseeable future. We know of no path that will take us from where we are now … to the Singularity” (Wooldridge, 2020, p. 242). This is an important topic, which deserves serious thought – but not at the expense of drawing attention away from the current and real uses of AI that hold huge potential for enormous benefits to humanity, but also for harming it (Bender & Hanna, 2023). This discussion will therefore focus on what AI could currently, or soon be able to, do; but it is important to keep in mind that advancements can happen very quickly and potentially take the public, and regulators, by surprise. Regulations therefore need to be designed to be flexible and agile enough to respond quickly to developments in new tech as they occur (Hacker, Engel, & Mauer, 2023).

AI may hold huge potential and we are still in the very earliest days of coming to grips with its abilities and the ways in which it is going to disrupt. Many high

[3] Turing’s imitation game test requires a computer to convince a human interrogator -- through written responses to questioning -- that it is a human. The test therefore involves a computer using deception, and as such, it is not a practical or ethical use of AI in general. However, the fact that students have used AI to dishonestly write assessment papers (Cotton, Cotton, & Shipway, 2023), and companies have used AI to respond to customer queries (Bharadiya, 2023; Okuda & Shoda, 2018) -- and in at least some of those cases the recipients were unaware that the written responses were created by machines -- means that in some senses the Turing test has been passed.

profile tech entrepreneurs have simultaneously lauded the potential benefits and also grave risks of emerging AI. The head of OpenAI, Sam Altman, has called for “coordinated international regulation” of AI, which, given the “risks of human extinction ... should be a global priority alongside other societal-scale risks such as pandemics and nuclear war” (Toh & Seo, 2023). Elon Musk and Apple co-founder Steve Wozniak were among thousands who signed an open petition calling for a six-month pause in the training of AI systems to “develop safety protocols for advanced AI design that makes AI systems more accurate, safer, interpretable, trustworthy, and loyal” (Ortiz, 2023). Bill Gates has likened it to other moments of great breakthrough innovations such as cars, personal computers and the internet, likening our current situation to the early days after the invention of the car; “those uncertain times before speed limits and seat belts” (Gates, 2023). Gates is optimistic about the ways in which “AI is going to revolutionize our lives. It will help solve problems—in health, education, climate change, and more” and that the risks, while real, can be managed by “adapting old laws and adopting new ones—just as existing laws against fraud had to be tailored to the online world”(Gates, 2023). When considering the statements of individuals with enormous vested interests in the outcomes of AI and its regulation it is wise to be circumspect about the motives that may lie behind their statements; this is not to say, though, that tech entrepreneurs cannot have valid and valuable opinions on the topic. Worth considering are, on the other hand, opposite views that call for an attention to the actual problems and harms of AI in the present and not the ones in a far and imagined future. We refer the reader to (Acemoglu, 2021; Bender, Gebru, McMillan-Major, & Shmitchell, 2021; Crawford, 2021; Kirkpatrick, 2023; Soper, 2021) for more on the computational costs, the ethical implications, or the impact AI has on peoples’ lives and the planet *now*.

This paper continues as follows: the next section discusses the speed at which AI is developing and the “Grand Dream” of AI compared with the actual current and likely abilities of AI in the near future. We argue that debates around possible future developments in AI should not detract from discussion about the real and current uses of this technology which simultaneously promise to enhance and threaten to diminish human liberty, dignity and prosperity. Then, we examine some of the main ethical frameworks and governmental approaches to regulating AI, and conclude that, while some progress has been made, there is much more work to be done in filling the gaps of the actual implementation of those frameworks and approaches. For example, transparency and accountability will present serious challenges in a field where those who research and build AI-based systems, themselves sometimes do not know why or how they produce the results that they do. We address how to drive attention to the underpinning ethical frameworks, as well as deal with the vexed question of intercultural values in ethical debates. The paper concludes with some recommendations to improve the discussion and outcomes around the ethics and regulation of contemporary machine learning-based AI.

In our analysis, we draw on a combination of a legal doctrinal approach which examines legislation, regulations, codes, bills and so on, with a philosophical framework drawing from the field of ethics and intercultural communication theory. The methodology will draw from secondary sources, including academic journal articles, media articles, governmental reports, and grey literature. The paper makes an original contribution by highlighting gaps in the ethical approach to thinking about Artificial Intelligence and considering the intercultural aspects of AI ethics which are rarely discussed in debates about 'responsible AI.'

The Grand Dream of AI

As Wooldridge points out in his book, the "grand dream" of AI, simply stated, is "to build machines that are self-aware, conscious and autonomous in the same way that people like you and me are" (Wooldridge, 2020, p. 2). However, he is quick to point out that "it is fiercely contentious – there isn't even any consensus that this kind of AI is feasible, let alone desirable" (Wooldridge, 2020, p. 2). While there are those seeking to recreate a digital version of the human brain, like the Human Brain Project for example (Staughton, 2022), there are others who argue that it is simply impossible to recreate human consciousness due to the complexity of the brain architecture; the neuronal electromagnetic fields or 'NEMFs' in the human brain, it is argued, create "a profound and unassailable chasm between the mammalian brain and any digital computer" or "Turing Machine" (Cicurel & Nicolelis, 2015, p. 19).

When ChatGPT 4 was used to assess its "IQ" in March 2023, it reportedly scored 155 on verbal IQ which is "higher than 99.9% of the 2,450 test takers that make up the standardization sample" (Siegal, 2023). However, there are at least as many works refuting such studies as the number of studies that test how intelligent AI-based artefacts (see e.g. (Bishop, 2021; Chen, Zaharia, & Zou, 2023; Mitchell, 2023)). For some, AI may appear intelligent, but comparing the human brain with computers is an enormously complex area fraught with difficulty and well beyond the scope of this article. We refer the reader to (Fjelland, 2020; Landgrebe & Smith, 2023; Mitchell, 2021; Shevlin, Vold, Crosby, & Halina, 2019), for example, for comprehensive analyses on why there are no truly intelligent artefacts on the horizon yet and there probably never will be. Landgrebe and Smith, for instance, actually state and extensively demonstrate this impossibility mathematically, computationally, evolutionarily, and neurobiologically. Turing himself estimated the storage capacity of the human brain as between "10^{10} to 10^{15} binary digits" stating "I incline to the lower values and believe that only a very small fraction is used for the higher types of thinking" (Turing, 1950, p. 455). It must also be noted that AI is not infallible, far from it; examples of AI errors and 'hallucinations' are plentiful (Alkaissi & McFarlane, 2023; Salvagno, Taccone, & Gerli, 2023), and self-driving cars have caused, or at least failed to avoid, fatal crashes (Ergin, 2022; Nyholm, 2018). What can perhaps be agreed upon is that when it comes to mathematical and statistical calculations, computers have a clear advantage over a human competitor. Computers are able to store and access phenomenally vast amounts of data and perform calculations on that data at incredible speed. As Wooldridge points out, “computers are *fast*. Very, very, very fast … a reasonable desktop computer operating at full speed can carry out up to *100 billion instructions of the type listed above every second*. One hundred billion is approximately the number of stars in our galaxy … it would take you about 3,700 years to do what the computer does in just one second” (Wooldridge, 2020, p. 23). And this kind of laptop capacity is dwarfed by the world’s fastest supercomputers (Greene, 2023); not to mention quantum computers which are predicted to be “a million times faster” than current supercomputers (Aiswarya, 2023).

AI, with access to data across huge swathes of stored human knowledge and deployed using techniques like deep learning, whose performance depends not only on the quality but also on the amount of this data (the more, the better), may be able to identify patterns and make trans-disciplinary connections in a way, and at a *speed*, that no human could ever hope to do (Cao, 2023). This is where the huge potential for human advancement lies: it might be possible for AI to contribute to solving intractable problems like curing cancer, climate change, alleviating poverty, boosting food and energy production, reducing transport and logistical challenges, etc. Other possible benefits might include improving access to legal or medical advice (Armstrong, 2023); or automated customer service with human-like interaction (Eliot, 2023). Detecting patterns, and irregularities in those patterns, is where AI excels. This is why the potentials of AI in the field of radiology are promising, although the limitations might be hugely challenging to overcome (Codari et al., 2019; Shin, Han, Ryu, & Kim, 2023; Vasilev et al., 2023; Waller et al., 2022).

Detecting patterns in data, however, could also be very useful for identifying patterns in the stock market (Raju et al., 2023); surveilling the movements and communications of populations of individuals both online and offline (Raju et al., 2023); identifying the movements of military assets and prioritising targets (Brose, 2020; Lee, 2023; Malmio, 2023); even for monitoring the health, exercise and consumption habits of citizens (Ali et al., 2023); and any number of other uses that will be highly attractive to governments and regimes that tend towards authoritarianism. Indeed, anti-fascist approaches to AI is a growing area of scholarship (McQuillan, 2022). This is to say nothing of the potential advantages to criminal groups in creating new ways of hacking and scamming (Renaud, Warkentin, & Westerman, 2023); or students using it to cheat on assessments (Cotton et al., 2023). The ability for ‘semantic reconstruction of language from non-invasive brain recordings’ (Tang, LeBel, Jain, & Huth, 2023) – also known as ‘mind reading’ (Samuel, 2023) will surely be of equal interest to law enforcement as it is to marketing/PR practitioners – but raises enormous ethical concerns. Self-driving vessels or drone boats may change the face of both terrorism and offensive naval operations (Fenton, 2023). Religious chatbots speaking in the 'voice of God' have condoned violence – the potential use by terrorist groups to radicalise individuals is highly concerning (Nooreyezdan, 2023). On the other hand, the advantages of AI for legitimate law enforcement are significant (Rademacher, 2020). However, the ongoing debate over encrypted communication apps is demonstrative of the tension that can arise from technological advances and ethical issues like privacy and justice. If there is a technological fix that could reduce or solve crimes, particularly crimes against children and other vulnerable groups, is there not a strong ethical obligation to use it? (Milmo, 2022).

Ethical AI and Regulation

There has been much discussion about ethical, or trustworthy AI, and how it should be regulated, if at all (Koniakou, 2023; Stahl et al., 2023). Some real progress has been made in this area with developments like the EU's AI Act "the world's first comprehensive AI law" (EU, 2023a, 2023b), a US "Blueprint for an AI Bill of Rights" (US, 2023), and a UK strategy for a "pro-innovation approach" to regulating AI (UK, 2023), to name just three prominent examples among many others. Noting that "there are over 600 AI-related policy recommendations, guidelines or strategy reports" from governments, NGOs, and private companies, Jobin, Ienca and Yayena (2019, p. 389) point out "there is a global convergence around five ethical principles: Transparency, Justice and Fairness, Non-Maleficence, Responsibility, and Privacy". While this sounds impressive, other observers are far more sceptical about progress in ethical AI. Thomas Metzinger, a member of the EU High-Level Expert Group on AI (AI HLEG), argues:

> "The Trustworthy AI story is a marketing narrative invented by industry, a bedtime story for tomorrow's customers. The underlying guiding idea of a "trustworthy AI" is, first and foremost, conceptual nonsense. Machines are not trustworthy; only humans can be trustworthy (or untrustworthy). If, in the future, an untrustworthy corporation or government behaves unethically and possesses good, robust AI technology, this will enable more effective unethical behaviour. Hence the Trustworthy AI narrative is, in reality, about developing future markets and using ethics debates as elegant public decorations for a large-scale investment strategy." (Metzinger, 2019a)

In fact, if the companies that make AI are not ethical, the AI itself will not be ethical (Monett, 2023). In other words, there is no way to code moral values, ethics, cultural influences, traditions, nor human societies' history in machines, i.e. *to code* them in any meaningful nor complete way in silicon. How AI artefacts (systems, programs, apps, robots, etc.) behave will always strongly depend on the data with which it was trained, on who owns that data and makes or dictates the decisions, on the values and ethics of the AI creators, of the companies that deploy it, of the users that use it, as well as on the people behind the curtains that make it possible for AI to even work (Crawford, 2021; Sap et al., 2022). In light of the many ethical transgressions of big tech – the Facebook Cambridge Analytica scandal being one prominent example (Brown, 2020) – it would seem to be appropriate to proceed with caution and scepticism. Metzinger warned about "ethics-washing" and pointed out that there were "hardly any ethicists" in the EU AI HLEG, which consisted of "four ethicists alongside 48 non-ethicists – representatives from politics, universities, civil society, and above all industry. That's like trying to build a state-of-the-art, future-proof AI mainframe with 48 philosophers, one hacker and three computer scientists" (Metzinger, 2019b).

Whereas the philosophical field of ethics has been well developed over centuries and has much to offer (Singer, 1991) – including the approaches set out by the main schools of deontological, utilitarian, and virtue ethics, among others – there has been scant reference to any specific ethical theories in the many public debates, reports, and papers on the topic of ethical AI. This deficit clearly requires attention from experts who are familiar with the various systems of ethics and have much to contribute. Technologies which threaten the freedom, dignity, prosperity and very lives of large segments of society in the ways outlined above, draw out the deep and real differences in ethical approaches that focus on the consequences and social utility of actions (utilitarianism), as opposed to those that emphasise the categorical imperative to treat others as morally significant "ends" rather than as "means to ends" (Kant, 2003).

Nevertheless, the pressing nature of ordering an ethical and regulatory approach to AI is perhaps not seen as conducive to deep philosophical debates by policy makers who require clear and fast standards and outcomes. Of the governmental approaches mentioned above, the EU's AI Act is perhaps the most ambitious insofar as it proposes the banning of certain types of uses of AI such as those that use "cognitive behavioural manipulation of people or specific vulnerable groups," "social scoring," as well as "real-time and remote biometric identification systems, such as facial recognition" (EU, 2023b). Other AI applications will be categorised according to their "risk of harm to health and safety, or an adverse impact on fundamental rights, to the environment, or to democracy and the rule of law" (EU, 2023a) and may be allowed where the company is registered and certain safeguards and risk management are in place.

Figure 1 illustrates the EU's risk-based approach to regulation, identifying four levels of AI risk.

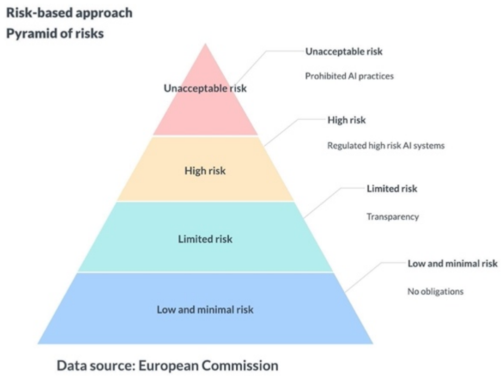

Figure 1: EU AI Act risk-based approach to regulating AI; created by authors based on (EU, 2023b)

Providers of "high-risk" AI systems – including biometric identification, law enforcement, education, and critical infrastructure management, among others – would be required to register their systems in an EU-wide database managed by the Commission before placing them on the market, and "would have to comply with a range of requirements particularly on risk management, testing, technical robustness, data training and data governance, transparency, human oversight, and cybersecurity (Articles 8 to 15)" (EU, 2023b, p. 5). Generative AI, like ChatGPT, categorised as "limited risk," would have to comply with transparency requirements; that is, disclosing that the content was AI-generated and preventing generation of illegal content, including summaries of copyrighted materials used in training the AI (EU, 2023b). On "deep fake" tech, the draft regulation points out the makers "shall disclose that the content has been artificially generated or manipulated," and also with regard to emotion or biometric recognition systems "shall inform in a timely, clear and intelligible manner of the operation of the system the natural persons exposed thereto and obtain their consent prior to the processing of their biometric and other personal data" ("EU AI Act," 2021). The EU approach is summarised thus:

> "Parliament's priority is to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly. AI systems should be overseen by people, rather than by automation, to prevent harmful outcomes. Parliament also wants to establish a technology-neutral, uniform definition for AI that could be applied to future AI systems." (EU, 2023b)

The US White House Blueprint for an AI Bill of Rights has "identified five principles that should guide the design, use, and deployment of automated systems to protect the American public in the age of artificial intelligence" (US, 2023, p. 3). Those principles are: safe and effective systems; algorithmic discrimination protections; data privacy; notice and explanation; human alternatives, consideration, and fallback (US, 2023).

The UK approach is slightly different insofar as it is presented as "A pro-innovation approach to AI regulation" which seeks "a common-sense, outcomes-oriented approach … to deliver better public services, high quality jobs and opportunities" to ensure that the UK will "become a science and technology superpower by 2030" and "the best place in the world to build, test and use AI technology" (UK, 2023, p. 1). Yet, the tone of the white paper from the outset is more bullish and optimistic about the potential benefits of AI, rather than being framed as a threat, which requires built-in protections or a Bill of Rights. This is not to say that safety and security have been ignored; the framework is underpinned by five principles to guide and inform responsible AI development (with echoes of the Bill of Rights and those outlined by Jobin et al. (2019), that is, safety, security and robustness, appropriate transparency and explainability, fairness, accountability and governance, contestability and redress). The UK however, seeking to remain "agile" and pro-innovation, "will not put these principles on a statutory footing initially … to avoid placing undue burdens on businesses" (UK, 2023, p. 6). The message to the international tech world, in a distinctly post-Brexit way, is clear: come to the UK to develop your innovative AI tech.

As Jobin and colleagues (2019) have pointed out, across the various approaches there are indeed recurring themes. In short, that AI should be safe, secure and do no harm – non-maleficence; it should not lead to or facilitate bias or discrimination against certain classes of people – like racial minorities or the poor; it should be transparent and accountable – that is, if a decision made by AI affects an individual's life, it must be able to find out how that decision was made and be able to challenge it; and privacy – an individual's private personal data should be protected.

Challenges to Ethical AI

While all of these principles are perfectly legitimate, it is unlikely that any of them can straightforwardly be guaranteed, despite what regulators would like to believe, or to have the public believe. Many of these issues are not unique to the AI context, and many of them continue to be a problem with current forms of digital processing, storage and cybersecurity. If they have not been solved in the context of 'traditional' computing yet, why should they be solvable in the discussion around AI?

Take privacy and data protection, for example; hardly a day goes by without a major news story of a hack resulting in the leaking of sensitive data from public and private organisations. Certain online media catalogue the daily hacks of various companies and organisations (Cybercrime, 2021). Attribution continues to be a major obstacle in traditional cybersecurity, that is, the ability to determine with certainty the real actor behind some kind of online crime or activity (Putrevu, Chunduri, Putrevu, & Shukla, 2023). Furthermore, large language models themselves like ChatGPT are prone to "jailbreaks" and other kinds of attacks (Shi, Liu, Zhou, & Sun, 2023; Zou, Wang, Kolter, & Fredrikson, 2023).

Explainability, interpretability or transparency may be a particularly difficult challenge since, as several reports have noted, in many cases the scientists who built AI-based applications "can't tell you how it works" (Hassenfeld, 2023). According to NYU's AI scientist Sam Bowman, "If we open up ChatGPT or a system like it and look inside, you just see millions of numbers flipping around a few hundred times a second, and we have no idea what any of it means" (Hassenfeld, 2023, p. 2). More concerning for regulators perhaps is "we don't have any good ideas yet about how to either technically control it or institutionally control it" (Hassenfeld, 2023, p. 37). The difficulty derives from two fundamentally different approaches to building AI: logic-based, symbolic AI and sub-symbolic, connectionistic AI (like neural network-based and other machine learning approaches). The former, as the name suggests, works from a set of explicit logical instructions to deductive outcomes. The latter takes an inductive, behavioural reinforcement approach where 'good' outcomes are rewarded and 'bad' outcomes are penalised until patterns emerge. A much needed third kind of reasoning, abduction-based, however, is almost absent from current AI applications.

As Bowman explains:

> "I think the key distinction is that with normal programs, with Microsoft Word, with Deep Blue [IBM's chess playing software], there's a pretty simple explanation of what it's doing. We can say, "Okay, this bit of the code inside Deep Blue is computing seven [chess] moves out into the future." … With these neural networks, there's no concise explanation. … All we can really say is just there are a bunch of little numbers and sometimes they go up and sometimes they go down. We don't have the concepts that map onto these neurons to really be able to say anything interesting about how they behave." (Hassenfeld, 2023, p. 29)

This is not to say that in the future there might not be ways of mapping the calculations of a neural net in a way that is interpretable for and by humans, but at present it remains "extremely, extremely hard" (Hassenfeld, 2023, p. 29). In this context, Metzinger's comments about 'ethical AI' being mere window dressing, 'public decorations' or 'bedtime stories,' start to take on a new urgency and relevance.

On the topic of non-maleficence, to return to Bill Gates' analogy to other great human technological breakthroughs like cars and the internet, if we had said to the inventor of the car or the internet before they had been released, "you must ensure that this invention does no harm," what would have been the result? Seemingly, AI developers are in a similarly difficult, if not impossible, position. They have a technology with enormous potential for good and harm, and yet it is simply impossible to give guarantees that it will be used exclusively for good and will not cause harm. Critical AI scholarship takes a closer look at these dilemmas and issues more formally (Birhane, Kasirzadeh, Leslie, & Wachter, 2023; Sætra, Coeckelbergh, & Danaher, 2022; Strümke, Slavkovik, & Madai, 2022).

What is notably absent from discussions around ethical AI, is the distribution of the economic benefits derived from AI. Let us be clear, if there are enormous profits to be made from AI, who will benefit most from those profits? Is there not an ethical imperative to ensure that those benefits are distributed fairly across society? One of the first statements on ethical AI, the Asilomar principles formulated in 2017, is a set of 23 principles that AI scientists and developers around the world were asked to sign up to (FLI, 2023). Included in the principles at number 15 is "the economic prosperity created by AI should be shared broadly, to benefit all humanity" (FLI, 2023). As Wooldridge warns, though:

> "It is hopelessly naïve to imagine that big businesses will do anything more than pay lip service to it. Big businesses are mainly investing in AI because they hope it will give them a competitive advantage that will deliver benefits to their shareholders, not because they want to benefit humanity." (Wooldridge, 2020, p. 254)

This is particularly relevant in the case of those whose data has been scraped without consent to train the AI models (Krotov & Johnson, 2023). It is an issue that is playing out in the Hollywood writers and actors' strike, and is an 'elephant in the room' for big tech and big business (Child, 2023). Ensuring the non-maleficence of AI is quite different from ensuring that powerful corporate elites or political establishments do not monopolise it for their own enrichment and privilege. The beginnings of this, one-AI-for-me-another-for-you trend, can already be seen insofar as the latest version of ChatGPT is no longer free.

This alludes to the issue of intercultural values and ethical AI. When regulators say "AI must be ethical," we may reasonably ask, whose ethics and for whom? Just as there are vast differences in cultural outlooks – an entire field of academic study has developed around intercultural communication (see (Hofstede, 2009), for example) – there are vastly different cultural approaches to ethics which link back to history, religion, geography, and deep attitudes, authority, family, nature, time and others. A culture which emphasises individualism, material success, and private property will have different attitudes to the distribution of benefits, than a culture that is collectivist and emphasises group harmony and mutual prosperity. For those reasons, it will be for individual states to analyse and decide for themselves how AI should be regulated in their jurisdictions. Having a universal ethics is not only utopic but also impossible.

If we examine the ethical approaches to AI and its regulation, as we have done with some ethical principles from the EU, for example, we can see that they focus, mainly, on what we call the outcomes of the ethical approach rather than the underlying rationale, the underpinning ethical infrastructure. Take transparency, for example: *why* does an individual have the right to know if they are dealing with, or affected by, an AI? In all the reports, debates and white papers, this is rarely, if ever, mentioned. To take another example, non-maleficence: *why* is it important that AI should do no harm? Couldn't one argue that in the pursuit of the many potential benefits of AI, it is legitimate and acceptable that *some* people will be harmed? One approach (consequentialism) emphasises the social utility or benefit of the many, over the rights of the few or the one. Another approach (Kantianism) treats individuals as *ends* in themselves which may never be sacrificed despite any social benefits because humans are morally significant subjects (and agents) whose actions have universal moral consequences.

This is why understanding the distinction between consequentialist and deontological approaches is key to understanding the whole framework of ethical systems that do not derive from religious dogma. It is a significant step forward in the advancement of understanding human relations, and at the same time it is largely absent from debates on the topic of ethical AI. Ethical context gives meaning to the entire debate around AI and its effects, and greater training in this area would be a major step forward both in the AI debate, in particular, and in human affairs, in general. Where regulation fails, only ethics can fill the vacuum.

CONCLUSION AND RECOMMENDATIONS

Clearly, AI is a technology with the *potential* for profound impacts on all aspects of human activity, from business to politics, over justice, education, and arts, to science, medicine, and virtually any other field of human endeavour. That is the hope at least. The challenge, as many academics, practitioners, observers, and some legislatures have realised, lies in regulating AI and ensuring that its use is ethical; that it is used for good rather than evil, to benefit humanity rather than destroy or damage it.

The following modest recommendations are made to that end:

a. Greater understanding and training of the underpinning ethical systems for regulators, but also for the practitioners developing AI, is needed so that when regulation fails, which it will, individuals and corporations are better placed to make ethical decisions without the need for a regulator looking over their shoulder.
b. It is difficult to predict exactly how capable AI is likely to become or what types of abilities it could have in the future. The blanket term ‘AI’ itself is vague; does it mean generative natural language processing? Machine learning? Robotics? Self-driving vehicles? Situational awareness and problem solving? Greater training and understanding is needed for regulators and the general public about what exactly are we talking about when we talk about ‘AI,’ AI literacy, in fact, is much broader a concept; it “should not be limited to learning about tools and technologies, but should also aim to equip providers and users with the notions and skills required to ensure compliance with and enforcement of [the EU AI Act] Regulation” (EU, 2023b).
c. The different sectors and strands of AI will require different solutions, policies and regulations. A sectoral approach which recognises these

differences is needed. Autonomous ships, for example, will be subject to a different legal regime than self-driving cars. This sector-specific knowledge and debate will be crucial at the stage of regulating specific sectors of industry. There is no point talking about regulating actions that a particular type of AI is not capable of. A chatbot like ChatGPT is not capable of driving a car, so discussions about generative language need not be concerned about things like 'STOP' signs.

REFERENCES

Acemoglu, D. (2021). Harms of AI. NBER Working Paper 29247. *SSRN Electronic Journal*.

Aiswarya, P. M. (2023). *Quantum Computing: Why is it Better Than Supercomputers?* Analytics Insight. https://www.analyticsinsight.net/quantum-computing-why-is-it-better-than-supercomputers/

Ali, O., Abdelbaki, W., Shrestha, A., Elbasi, E., Alryalat, M. A. A., & Dwivedi, Y. K. (2023). A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *Journal of Innovation & Knowledge, 8*(1), 100333. doi: 10.1016/j.jik.2023.100333

Alkaissi, H., & McFarlane, S. I. J. C. (2023). Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*, 19;15(2):e35179. doi: 10.7759/cureus.35179

Armstrong, K. (2023). *ChatGPT: US lawyer admits using AI for case research*. BBC. https://www.bbc.co.uk/news/world-us-canada-65735769

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada. doi: 10.1145/3442188.3445922

Bharadiya, J. P. (2023). Machine Learning and AI in Business Intelligence: Trends and Opportunities. *International Journal of Computer (IJC), 48*(1), 123-134.

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the age of large language models. *Nature Reviews Physics, 5*(5), 277-280. doi: 10.1038/s42254-023-00581-4

Bishop, J. M. (2021). Artificial Intelligence Is Stupid and Causal Reasoning Will Not Fix It. *Frontiers in Psychology, 11*. doi: 10.3389/fpsyg.2020.513474

Brose, C. (2020). *The Kill Chain: Defending America in the Future of High-Tech Warfare*: Hachette Books.

Brown, A. J. (2020). "Should I Stay or Should I Leave?": Exploring (Dis)continued Facebook Use After the Cambridge Analytica Scandal. *Social Media + Society, 6*(1), 2056305120913884. doi: 10.1177/2056305120913884

Cao, L. (2023). Trans-AI/DS: transformative, transdisciplinary and translational artificial intelligence and data science. *International Journal of Data Science and Analytics, 15*(2), 119-132. doi: 10.1007/s41060-023-00383-y

Cath, C., & Jansen, F. (2021). Dutch Comfort: The limits of AI governance through municipal registers. *Techné: Research in Philosophy and Technology, 26*(3), 395-412. doi: 10.5840/techne202323172

Chalmers, D. J. (2010). The Singularity: a philosophical analysis. *Journal of Consciousness Studies*, 17:7-65, 2010.

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* ArXiv, /abs/2307.09009.

Child, B. (2023). *AI is coming for Hollywood scriptwriters – this is how they are going to do it*. The Guardian. https://www.theguardian.com/film/2023/may/12/ai-artificial-intelligence-generating-screenplays

Cicurel, R., & Nicolelis, M. A. L. (2015). *The Relativistic Brain: How it works and why it cannot be simulated by a Turing Machine*. Natal: Kios Press.

Codari, M., Melazzini, L., Morozov, S. P., van Kuijk, C. C., Sconfienza, L. M., Sardanelli, F., & European Society of, R. (2019). Impact of artificial intelligence on radiology: A EuroAIM survey among members of the European Society of Radiology. *Insights into Imaging, 10*(1), 105. doi: 10.1186/s13244-019-0798-3

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 1-12. doi: 10.1080/14703297.2023.2190148

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*: Yale University Press.

Cybercrime. (2021, 2021-01-28). Who's Hacked? Latest Data Breaches And Cyberattacks. Cybercrime Magazine. https://cybersecurityventures.com/intrusion-daily-cyber-threat-alert/

Eliot, L. (2023). *Here's How ChatGPT Can Either Enhance Customer Service Or Brutally Undermine The Customer Experience, Warns AI Ethics And AI Law.* Forbes. https://www.forbes.com/sites/lanceeliot/2023/04/21/heres-how-chatgpt-can-either-enhance-customer-service-or-brutally-undermine-the-customer-experience-warns-ai-ethics-and-ai-law/

Ergin, U. (2022). One of the First Fatalities of a Self-Driving Car: Root Cause Analysis of the 2016 Tesla Model S 70D Crash. *Trafik ve Ulaşım Araştırmaları Dergisi, 5*(1), 83-97. doi: 10.38002/tuad.1084567

EU (2023a). Artificial Intelligence Act. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. In *(COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*. EU: EU Parliament.

EU (2023b). *EU AI Act: first regulation on artificial intelligence*. European Parliament. European Parliament News. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

EU AI Act (2021). EUR-Lex - 52021PC0206 - EN - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

Fenton, A. J. (2023). *Ukraine: how uncrewed boats are changing the way wars are fought at sea*. The Conversation. https://theconversation.com/ukraine-how-uncrewed-boats-are-changing-the-way-wars-are-fought-at-sea-201606

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*(1), 10. doi: 10.1057/s41599-020-0494-4

FLI (2023). *AI Principles*. Future of Life Institute. https://futureoflife.org/open-letter/ai-principles/

Gates, B. (2023). *How to manage risks of AI*. Gates Notes. https://www.linkedin.com/pulse/how-manage-risks-ai-bill-gates/

Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers*, 6, 31-88. doi: 10.1016/S0065-2458(08)60418-0.

Greene, T. (2023). *Frontier still reigns as world's fastest supercomputer*. Network World. https://www.networkworld.com/article/3697308/frontier-still-reigns-as-the-world-s-fastest-supercomputer.html

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'23 (1112–1123). doi: 10.1145/3593013.3594067.

Hassenfeld, N. (2023). *Even the scientists who build AI can't tell you how it works*. Vox. https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, 2(1). https://doi.org/10.9707/2307-0919.1014

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. doi: 10.1038/s42256-019-0088-2

Kant, I. (2003). *Critique of pure reason* (T. M. Weigelt, Trans.): Penguin Classics.

Kekatos, M. (2023). *How artificial intelligence is being used to detect,* *treat cancer -- and the potential risks for patients*. abc News. https://abcnews.go.com/Health/ai-detect-treat-cancer-potential-risks-patients/story?id=101431628

Kirkpatrick, K. (2023). The Carbon Footprint of Artificial Intelligence. *Communications of the ACM, 66*(8), 17-19.

Koniakou, V. (2023). From the "rush to ethics" to the "race for governance" in Artificial Intelligence. *Information Systems Frontiers, 25*(1), 71-102. doi: 10.1007/s10796-022-10300-6

Krotov, V., & Johnson, L. (2023). Big web data: Challenges related to data, technology, legality, and ethics. *Business Horizons, 66*(4), 481-491. doi:https://doi.org/10.1016/j.bushor.2022.10.001

Landgrebe, J., & Smith, B. (2023). *Why machines will never rule the world: artificial intelligence without fear*. Routledge Taylor & Francis.

Lee, J. (2023). *"Overtaking on the Curve?" Defense AI in China*. Defense AI Observatory. https://defenseai.eu/wp-content/uploads/2023/04/daio_study2313_overtaking_on_the_curve_john_lee.pdf

Malmio, I. (2023). Ethics as an enabler and a constraint – Narratives on technology development and artificial intelligence in military affairs through the case of Project Maven. *Technology in Society, 72*, 102193. doi: 10.1016/j.techsoc.2022.102193

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. (2023). The AI Index 2023 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, USA. https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

McQuillan, D. (2022). *Resisting AI: An anti-fascist approach to Artificial Intelligence*. UK: Bristol University Press.

Metzinger, T. (2019a). *EU guidelines: Ethics washing made in Europe*. Tagesspiegel. https://www.tagesspiegel.de/politik/ethics-washing-made-in-europe-5937028.html

Metzinger, T. (2019b). *Expert commentary: "Ethics washing" made in Europe*. Business & Human Rights Resource Centre. Business and Human Rights Resource Centre. https://www.business-humanrights.org/en/latest-news/expert-commentary-ethics-washing-made-in-europe/

Milmo, D. (2022). *NCA says end-to-end encryption poses challenge for law enforcers on child abuse*. The Guardian. https://www.theguardian.com/uk-news/2022/jan/22/nca-says-end-to-end-encryption-challenge-law-enforcers

Mitchell, M. (2021). Why AI is harder than we think. In *Proceedings of the Genetic and Evolutionary Computation Conference*, Lille, France. doi: 10.1145/3449639.3465421

Mitchell, M. (2023). *Did ChatGPT really pass graduate-level exams*. Substack AI: A guide for thinking humans. https://aiguide.substack.com/p/did-chatgpt-really-pass-graduate-baa

Monett, D. (2023). *Rethinking how to deal with contemporary AI: Ethics is not a ballast but a need.* Talk at the Centre of Business in Society, Coventry University, UK.

Nooreyezdan, N. (2023). *India's religious AI chatbots are speaking in the voice of god — and condoning violence*. rest of world. https://restofworld.org/2023/chatgpt-religious-chatbots-india-gitagpt-krishna/

Nyholm, S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *J Philosophy Compass, 13*(7), e12507.

Okuda, T., & Shoda, S. (2018). AI-based chatbot service for financial industry. *Fujitsu Scientific and Technical Journal, 54*(2), 4-8.

Ortiz, S. (2023). *Musk, Wozniak, and other tech leaders sign petition to halt further AI developments*. ZD Net. https://www.zdnet.com/article/musk-wozniak-and-other-tech-leaders-sign-petition-to-halt-ai-developments/

Putrevu, V. S. C., Chunduri, H., Putrevu, M. A., & Shukla, S. (2023). A Framework for Advanced Persistent Threat Attribution using Zachman Ontology. In *Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference*, Stavanger, Norway. doi: 10.1145/3590777.3590783

Rademacher, T. (2020). Artificial Intelligence and Law Enforcement. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating Artificial Intelligence* (pp. 225-254). Cham: Springer International Publishing.

Raju, S. S., Srikanth, M., Guravaiah, K., Pandiyaan, P., Teja, B., & Tarun, K. S. (2023, 11-12 Feb. 2023). A Three-Dimensional Approach for Stock Prediction Using AI/ML Algorithms: A Review & Comparison. In *Proceedings of the 2023 4th International Conference on Innovative Trends in Information Technology* (ICITIIT), Kottayam, India, 2023, pp. 1-6, doi: 10.1109/ICITIIT57246.2023.10068584.

Renaud, K., Warkentin, M., & Westerman, G. (2023). From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI. *MIT Sloan Management Review*, Reprint No. 64428. https://sloanreview.mit.edu/article/from-chatgpt-to-hackgpt-meeting-the-cybersecurity-threat-of-generative-ai/

Sætra, H. S., Coeckelbergh, M., & Danaher, J. (2022). The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. *AI and Ethics, 2*(1), 15-27. doi: 10.1007/s43681-021-00123-7

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial intelligence hallucinations. *Critical Care, 27*(1), 1-2.

Samuel, S. (2023). *Mind-reading technology has arrived*. Vox. https://www.vox.com/future-perfect/2023/5/4/23708162/neurotechnology-mind-reading-brain-neuralink-brain-computer-interface

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., & Smith, N. A. (2022). *Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection*. ArXiv, arXiv:2111.07997.

Shevlin, H., Vold, K., Crosby, M., & Halina, M. (2019). The limits of machine intelligence. *EMBO reports, 20*(10), e49177. doi: 10.15252/embr.201949177

Shi, J., Liu, Y., Zhou, P., & Sun, L. (2023). *BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT*. ArXiv, https://arxiv.org/abs/2304.12298v1. doi: 10.48550/arXiv.2304.12298

Shin, H. J., Han, K., Ryu, L., & Kim, E.-K. (2023). The impact of artificial intelligence on the reading times of radiologists for chest radiographs. *npj Digital Medicine, 6*(1), 82. doi: 10.1038/s41746-023-00829-4

Siegal, J. (2023). *ChatGPT took an IQ test, and its score was sky-high*. BGR. https://bgr.com/tech/chatgpt-took-an-iq-test-and-its-score-was-sky-high/

Singer, P. (1991). *A Companion to Ethics*. Wiley-Blackwell.

Smith, T. (2023). *Google Bard's New Visual Feature is a Game Changer*. Medium. https://medium.com/the-generator/google-bards-new-visual-feature-is-a-game-changer-b0fc28256ce4

Soper, S. (2021). *Fired by bot at Amazon:'It's you against the machine'*. Bloomberg, June, 28. https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z., & Wright, D. (2023). A systematic review of artificial intelligence impact assessments. *Artificial Intelligence Review*. doi: 10.1007/s10462-023-10420-8

Staughton, J. (2022). The Human Brain Vs. *Supercomputers... Which One Wins?* Science ABC. https://www.scienceabc.com/humans/the-human-brain-vs-supercomputers-which-one-wins.html

Strümke, I., Slavkovik, M., & Madai, V. I. (2022). The social dilemma in artificial intelligence development and why we have to solve it. *AI and Ethics, 2*(4), 655-665. doi: 10.1007/s43681-021-00120-w

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience, 26*(5), 858-866. doi: 10.1038/s41593-023-01304-9

Toh, M., & Seo, Y. (2023). *OpenAI CEO calls for global cooperation to regulate AI*. CNN Business.

https://www.cnn.com/2023/06/09/tech/korea-altman-chatgpt-ai-regulation-intl-hnk/index.html

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433-460.

UK (2023). *Policy paper AI regulation: a pro-innovation approach*. UK. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

US (2023). Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People. The White House. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

van Maanen, G. (2022). AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. *Digital Society, 1*(2), 9. doi: 10.1007/s44206-022-00013-3

Vasilev, Y., Vladzymyrskyy, A., Omelyanskaya, O., Blokhin, I., Kirpichev, Y., & Arzamasov, K. (2023). AI-Based CXR First Reading: Current Limitations to Ensure Practical Value. *Diagnostics, 13*(8), 1430.

Waller, J., O'Connor, A., Raafat, E., Amireh, A., Dempsey, J., Martin, C., & Umair, M. (2022). Applications and challenges of artificial intelligence in diagnostic and interventional radiology. *Polish Journal of Radiology, 87*(1), 113-117. doi: 10.5114/pjr.2022.113531

Wooldridge, M. (2020). *The Road to Conscious Machines: the story of AI*. UK: Penguin Random House.

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and Transferable Adversarial Attacks on Aligned Language Models. ArXiv, https://arxiv.org/abs/2307.15043v1. doi: 10.48550/arXiv.2307.15043