The Impact of Artificial Intelligence (AI) Developments on Culture and Society: Regulation, Control and Alignment

Terry Hyland[1]

1 Free University of Ireland

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

The manic flurry of activity following the recent introduction of Open AI's ChatGPT-4, Google's Bard and other similar advanced Large Language Models (LLMs) has tended to generate rather more heat than light in both popular and academic discourse about the main implications of the new applications. Debate has ranged from doom-laden apocalyptic warnings to hyperbolic accounts of how AI will revolutionise and enhance all aspects of human activity. Attempting to steer a middle way between these extremes, this article concentrates on the key issues of regulation, control and alignment of the new systems since these are the areas that are likely to be of the first importance in informing and influencing the ways in which AI impacts all aspects of our lives. The key themes examined here build on my previous article on the educational implications of AI which was published in Qeios (Hyland, 2023a), and there are some overlaps with this piece and the current discussion so as to provide a background for those who may not have read the original.

Professor Terry Hyland

Free University of Ireland, Dublin 7 –hylandterry@ymail.com

ORCID iD: 0000-0001-8539-8211

Keywords: Artificial intelligence (AI), AI regulation, AI control, AI/human value alignment.

Introduction

Although artificial intelligence (AI) systems have been with us for decades – in phones, cars, banking, medicine, and the like – it was the appearance of the publicly accessible Chat GPT and similar applications and tools in November 2022 that has stimulated such intense and unrelenting academic and popular interest. Much of the interest and debate has been located within academic disciplines since, given their natural responsibilities for educational development within their various fields (Hyland, 2023), the AI applications have had a direct impact on research, teaching and learning at all levels. The PubMed site records over 720,000 articles on AI applications over the last year or so (PubMed, 2024) and the more generalised ResearchGate website includes 16,200 relevant studies (ResearchGate, 2024). Debate tends to be polarised between accounts pointing to the dangers and existential risks associated with the new systems and those explaining the tremendous benefits of AI in all spheres of human activity.

Renaud Foucart (2023) offers a representative illustration in his comment that:

> *AI is expected to affect every aspect of our lives – from healthcare, to education, to what we look at and listen to, and even how how well we write. But AI also generates a lot of fear, often revolving around a god-like computer*

> *becoming smarter than us, or the risk that a machine tasked with an innocuous task may inadvertently destroy humanity. More pragmatically, people often wonder if AI will make them redundant (p.1)*

Many of the concerns about the new AI tools have been expressed by educators who fear that teaching and learning will be damaged by the easy access to the large language models (LLMs) like ChatGPT which can write essays and answer assignment questions in a matter of minutes. As Will Douglas Heaven (2023) observed in the MIT Technology Review:

> *Just days after OpenAI dropped ChatGPT in late November 2022, the chatbot was widely denounced as a free essay-writing, test-taking tool that made it laughably easy to cheat on assignments. Los Angeles Unified, the second-largest school district in the US, immediately blocked access to OpenAI's website from its schools' network. Others soon joined. By January, school districts across the English-speaking world had started banning the software, from Washington, New York, Alabama, and Virginia in the United States to Queensland and New South Wales in Australia. Several leading universities in the UK, including Imperial College London and the University of Cambridge, issued statements that warned students against using ChatGPT to cheat (p.2).*

On a a more general level, an open letter signed by, amongst other leading figures in the digital technology world, Elon Musk and co-founder of Apple, Steve Wozniak, called for a pause to current machine learning AI developments until the wider implications are evaluated carefully. The letter published by the *Future of Life Institute* (2023) warns that:

> *AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs**...** Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable...Therefore, we call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4**.** This pause should be public and verifiable, and include all key actors. If such a pause cannot be enacted quickly, governments should step in and institute a moratorium. (p.1)*

In a similar vein – in an announcement which stunned the AI tech world – Geoffrey Hinton, the so-called "godfather of AI" quit his job at Google warning of the dangers of unregulated and unsupervised tends in the field. Among his comments he observed that:

> *Some of the dangers of AI chatbots were "quite scary", warning they could become more intelligent than humans and could be exploited by "bad actors". "It's able to produce lots of text automatically so you can get lots of very effective spambots. It will allow authoritarian leaders to manipulate their electorates, things like that. But, he added, he was also concerned about the existential risk of what happens when these things get more intelligent than us. I've come to the conclusion that the kind of intelligence we're developing is very different from the intelligence we have. So it's as if you had 10,000 people and whenever one person learned something, everybody automatically knew it. And that's how these chatbots can know so much more than any one person. (Taylor &*

> *Hern, 2023)*

The most recent grim forecast was issued by ex-Google CEO, Eric Schmidt, who declared that there aren't enough guardrails to stop the technology from doing catastrophic harm. He warned that ‘After Nagasaki and Hiroshima, it took 18 years to get to a treaty over test bans and things like that...We don't have that kind of time today." (Al-Sibai, 2023).

AI Negatives: Risks and Dangers

Principal areas of concerns for those warning of the potential threats posed by AI include use of the new systems in defence and warfare, in the workplace, in education and, more generally, as a means of control, surveillance and discrimination by the police, the judiciary, and recruitment personnel. In terms of AI implications for jobs and the economy the chief concern in this area is that the new applications will replace many jobs currently performed by humans thus leading to mass redundancy and unemployment. British Telecom announced recently that it would be cutting its workforce by 55,000 with 11,000 of these jobs replaced by AI (BBC News, 2023). On a more dramatic scale, a recent report by Goldman Sachs predicted that around 300 million jobs would in future be lost or degraded by forms of AI (Kelly, 2023).

Though there may be disagreements about degrees of control and regulation of AI uses, few commentators disagree about the control of AI use in defence systems and warfare. Several examples of near nuclear disasters avoided by humans are on record – a recent film was made about the most famous one, the Russian, Stanislav Petrov (Myre, 2018) – and given the range of drones, tanks, fighter planes and missiles currently operating with AI components, the control and regulation of such applications is vital. Both military bodies and tech experts are now fully aware of the dangers here, and there are urgent calls in the US for ‘Congress to put guardrails in place to ensure [AI] is not misused’ (Khalil, 2023). Appropriate models of regulation and alignment with human values will be discussed later in terms of general implications of the new technology in all spheres of application.

The impact of AI on general online activity is causing considerable anxiety given the potential for creating deepfake misinformation for political or fraudulent purposes, and there is also the danger of discriminatory bias being proliferated through machine learning applications. As Nicoletti & Bass (2023) have commented:

> *Some experts in generative AI predict that as much as 90% of content on the internet could be artificially generated within a few years. As these tools proliferate, the biases they reflect aren’t just further perpetuating stereotypes that threaten to stall progress toward greater equality in representation — they could also result in unfair treatment. Take policing, for example. Using biased text-to-image AI to create sketches of suspected offenders could lead to wrongful convictions (p.2).*

In all such spheres in which decisions and judgements are made, the control and alignment principles recommended by Nick Bostrom (2014, 2023) and others need to be foregounded as an absolute priority in AI developments.

AI Positives: Benefits and Assets

Against the negative prognostications for AI there are many positive voices and visions predicting the revolutionary benefits to be gained in areas such as work, medicine, and education. In the sphere of work, a report by the *World Economic Forum* (2023), for example, makes the following important points about what commentators are calling the 'fourth industrial revolution':

- Around 40% of all working hours could be impacted by AI large language models (LLMs) such as ChatGPT-4, says a report from Accenture.
- Many clerical or secretarial roles are seen as likely to decline quickly because of AI, according to the World Economic Forum's Future of Jobs Report 2023.
- But roles for AI and machine learning specialists, data analysts and scientists, and digital transformation specialists are expected to grow rapidly, the report adds.
- Reskilling people to use AI effectively will be the key to companies being able to use the technology successfully, says Accenture.

The key message here is that:

> *Success with generative AI requires an equal attention on people and training as it does on technology...This means both building talent in technical competencies like AI engineering and enterprise architecture, and training people across the organization to work effectively with AI-infused processes (p.1).*

In a similar vein, Jonathan Aitken (2023) reminds us that:

> *The development of technology and its associated impact on job security has been a recurring theme since the industrial revolution. Where mechanisation was once the cause of anxiety about job losses, today it is more capable AI algorithms. But for many or most categories of job, retaining humans will remain vital for the foreseeable future.*

He goes on to suggest that:

> *This means that, as workers, we need to look to harness the capability of AI systems and use them to their full potential. This means always questioning what we receive from them, rather than just trusting their output blindly...If we apply a sceptical mindset to how we use this new tool, we'll maximise its capability while simultaneously growing the workforce – as we've seen through all the previous industrial revolutions (pp.1-2).*

In education, the enthusiasm for the AI revolution – emphasising the promise over the potential threats – is even more in evidence. Will Douglas Heaven (2023), for example, comments that initially educators were worried that 'ChatGPT would

undermine the way we test what students have learned, a cornerstone of education'. However, it seems that many teachers have now adapted to the new applications and discovered some positive ways of working with them. As he comments:

> *Far from being just a dream machine for cheaters, many teachers now believe, ChatGPT could actually help make education better. Advanced chatbots could be used as powerful classroom aids that make lessons more interactive, teach students media literacy, generate personalized lesson plans, save teachers time on admin, and more (pp.1-2).*

Similarly, Karen Lancaster (2023) urges university lecturers to embrace AI applications such as ChatGPT working with students to eliminate errors and achieve the best results. She concludes her plea for a working partnership between AI tools and academia by observing that:

> *if universities accept the use of AI software for essay-writing, they should increase the expected level of scholarship accordingly, in the same way that maths tests for people with calculators should demand a higher level of aptitude than maths tests for people without calculators. But simply behaving as if the technology doesn't exist, or decreeing that its use amounts to misconduct, is probably not a prudent way forward (p.4).*

Similar arguments have been advanced by Claire Chen (2023) of Stanford University who reviews a wide range of insights which can be gained from working creatively with AI tools in ways which can enhance teaching and learning in many fields of learning. It is well to acknowedge all the increments in learning and development – in just about every sphere of activity – that have been made in recent decades through AI tools. The fact that humans have improved at the fiendishly complex game of Go since DeepMind's AlphaGo finally defeated the world's best players (Rosenblum, 2023) is just one small, perhaps emblematic, indication of what educators and policymakers might gain by working in partnership with AI applications.

The use of AI in health and social care is another sphere which hold out exciting and valuable prospects for progress and the enhancement of provision. Liz Kwo (2021) has outlined a wide range of advances and gains made through the introduction of AI tools in healthcare settings – from drug discovery to rapid diagnosis – and concludes her review of the field by observing that by:

> *improving workflows and operations, assisting medical and nonmedical staff with repetitive tasks, supporting users in finding faster answers to inquiries, and developing innovative treatments and therapies, patients, payers, researchers and clinicians can all benefit from the use of AI in healthcare (p.1).*

Just recently, it was announced that AI had been used 'to discover abaucin, an effective drug against A baumannii, bacteria that can cause dangerous infections' (Yang, 2023), and the range of opportunities for improving all aspects of medical treatment through AI expands exponentially (Davenport, 2019).

AI Regulation

The exponential rate of AI development in recent months has been matched by the hectic scramble by nations and organisations to exert forms of regulation as a means of protection against potential threats. The White House has issued a *Blueprint for an AI Bill of Rights*(2023) which includes the following five fundamental principles to guide future implementation and use of of the new AI applications (pp.3-6):

- You should not face discrimination by algorithms and systems should be used and designed in an equitable way
- You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used
- You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you
- You should be able to opt out, where appropriate, and you should be protected from unsafe or ineffective systems
- You should have access to a person who can quickly consider and remedy problems you encounter

A similar set of policy principles have been proposed recently by the UK Government which have been described as a 'pro-innovation approach' designed to ensure that AI regulation does not interfere with investment in the new technologies. Albert Sanchez-Graells (2023) has pointed out that the 'plans have been criticised for being too lax, already outdated, and lacking in meaningful detail' (p.1). He suggests that the policy proposals have more to do with ensuring a post-Brexit AI technology market for Britain rather than protecting the public from potential harm, and concludes that:

Only by implementing effective protections and showing strong and decisive action domestically can the UK government hope to build the credibility needed to lead international efforts of AI regulation (ibid., p.2).

Such international efforts have been realised recently in a spate of conferences discussing the principal concerns and issues. This activity has resulted in the *European Artificial Intelligence Act* which includes the following provisions:

- Safeguards agreed on general purpose artificial intelligence
- Limitation for the of use biometric identification systems by law enforcement
- Bans on social scoring and AI used to manipulate or exploit user vulnerabilities
- Right of consumers to launch complaints and receive meaningful explanations
- Fines ranging from 35 million euro or 7% of global turnover to 7.5 million or 1.5% of turnover

The regulations are intended to ensure 'that fundamental rights, democracy, the rule of law and environmental sustainability are protected from high risk AI, while boosting innovation and making Europe a leader in the field. The rules establish obligations for AI based on its potential risks and level of impact' (European Parliament, 2023).

Similar provisions are likely to apply on a global level, and Open AI's CEO, Sam Altman, has given a cautious welcome to to such regulations, suggesting that this is what the new tech companies have been asking for. In an interview with

Azheem Azhar, Altman claimed that – after their own campaigning efforts to establish regulatory frameworks with government bodies – the corporation now has robust and transparent protocols and safety regulations in place. ChatGPT-4 is now apparently being "trained" to ensure alignment with human values and interests. Pending this exercise, work on ChatGPT-5 has been paused. (Bloomberg,2023. https://youtu.be/w5nEf-HahZM?si=cftdVwDVOxAd60sf).

AI Control: Internal

It is one thing to establish regulations to cover AI use and development, quite another to ensure that the necessary control mechanisms are built into the new technology. Bostrom's work on *Superintelligence* (2014) has established the definitive blueprint for progress in this crucial field. He initially outlines a range of 'capability control methods' which are designed to 'prevent undesirable outcomes by limiting what the superintelligence can do'. This is further explained in terms of:

> *placing the superintelligence in an environment in which it is unable to cause harm (boxing methods) or in which there are strongly convergent instrumental reasons not to engage in harmful behaviour (incentive methods). It might involve limiting the internal capacities of the superintelligence (stunting). In addition, capability control methods might involve the use of mechanisms to automatically detect and react to various kinds of containment failure or attempted transgression (tripwires). (2014, pp.157-8).*

Bostrom goes on to suggest that control methods need to be matched to specific AI tools and recommends international collaboration in the process of planning and development. However, he states quite frankly that 'capability control is, at best, a temporary and auxiliary measure' and that long-term safeguards must tackle the 'value-loading problem' (ibid., p.226). This problem – also known as the alignment problem of how to align the algorithmic computational power of AI with human needs and interests – is a difficult one to crack and and is an ongoing project. Bostrom recommends a 'common good principle' for all developments in this field; this is that

> *Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals (2014, p. 312).*

However, this leaves the philosophical problem (discussed below) of identifying the universal ethical ideals, in addition to the difficult technical problem of building such ideals into machine intelligence since, as Bostrom admits, 'it is not yet known how to use the value learning approach to install plausible human values' into AI (ibid., p.241).

Max Tegmark (2023) has suggested recently that the so-called alignment training currently operating on the advanced AI tools really amounts to training the applications what not to say rather than what not to do. This may be sufficient for the LLMs but would be clearly not nearly enough for AI applications in transport, medicine and defence fields. In these areas Tegmark claims that – until the goal-alignment and ethical problems of AI development are solved – failsafe mechanisms based on proof specifications need to be built into all new applications. Through the use of built-in algorithmic proof

checkers, therefore, we can approximate to the evolution of AI tools which provide greater protection from possible harms (ibid.)

AI Control: External

In addition to the internal mechanisms outlined above there will be a need to ensure that the regulatory frameworks for controlling AI established in the US and Europe are adhered to by nation states, corporations and independent developers. Open AI claims to have conducted extensive research on the control problem and its official position outlines a range of safeguarding provisions which:

> *include conducting pre-deployment risk assessments; external scrutiny of model behavior; using risk assessments to inform deployment decisions; and monitoring and responding to new information about model capabilities and uses post-deployment (Open AI, 2023, p.1).*

However, given the potential existential threats posed by AI developments – described in graphic detail by, for example, Toby Ord (2020) and Max Tegmark (2017) – it will be necessary to establish forms of external surveillance to ensure that any internal control mechanisms are closely monitored and enforced. After all, such mechanisms are routine and standard in areas such as medicine, social care, health and safety, environmental protection, and so on. An example taken from my own country of Ireland relating to the Health and Safety Authority framework of regulatory bodies might serve as a case in point. Listed below are the various bodies charged with regulating health, social care and related fields throughout Ireland (HSA, 2023).

HEALTH AND SAFETY AUTHORITY – IRELAND

Regulators in Healthcare

Regulator of Occupational Safety and Health

- Health and Safety Authority

Regulators of Services

- Health Information and Quality Authority
- Mental Health Commission
- Tusla - Child and Family Agency

Regulators of Professionals

- Nursing and Midwifery Board of Ireland
- Dental Council
- CORU - Regulating Health & Social Care Professionals
- Medical Council of Ireland
- Pharmaceutical Society of Ireland (PSI)
- Pre-Hospital Emergency Care Council (PHECC)

Regulators of Products

- Food Safety Authority of Ireland (FSAI)
- Health Products Regulatory Authority (HPRA)
- Environmental Protection Agency (EPA)

All of the above organisations and bodies would engage in regular inspections to ensure compliance with standards and, of course, as in all developed countries, there is a police force and a tax enforcement system to provide additional oversight and supervision. Similar regulatory frameworks would apply to factories, shops, energy and water providers and all establishments offering a service to the public. Such arrangements are standard and mainstream in most countries and a similar external regulatory and enforcement framework could conceivably be applied to AI development given the political will and necessary resources.

In addition to all such safeguards and regulatory frameworks there is the public voice, whether this is expressed through academic discourse referred to in this article or in popular journalism. Open AI is ostensibly a non-profit company though its work relies heavily on its multi-billion dollar partnership with Microsoft which is the most successful corporation in American business history. As a corporation, Microsoft has the primary aim of generating profit for its shareholders, hence the uneasy relationship with the allegedly non-profit Open AI which recently sacked its CEO, Altman, then rehired him within a day after he was said to have moved to Microsoft. The dispute within AI – according to Jason Hassett (2024) – may have been partly about disagreements over technical developments but also the idea of building a for-profit stream beneath the umbrella of an ostensibly non-profit scientific research company. Whatever the real story, such instability and

uncertainty seems to call for rigorous external oversight to protect the public.

In relation to public protection, it is also worth mentioning that corporations are legal persons (as noted later) and can be held accountable under civil and criminal law for their activities. Organisations like Facebook (now Meta) may at times seem to be a law unto themselves, though it is worth remembering that Facebook has faced numerous lawsuits – having to pay out millions in compensation – for the invasion of privacy, and in recent months 'dozens of states are suing Meta, alleging the tech giant has deliberately engineered its social media platforms Instagram and Facebook to be addictive to children and teens' (Gibson, 2023). Just recently it was announced that 'OpenAI and its close collaborator and investor, Microsoft' was being sued by the New York Times 'for allegedly violating copyright law by training generative AI models on the Times' content' (Wiggers, 2024). It is highly likely that such lawsuits will increase exponentially as AI impacts more and more spheres of human activity.

AI Values Alignment

Although most AI commentators would wish to endorse Bostrom's 'common good principles' for future superintelligent machine learning, this still leaves the problem of just what principles to incorporate in new technology. Tegmark (2017) discusses this issue at length and – in addition to examining a range of traditional ethical theories in the philosophical tradition – refers to Isaac Asimov's famous 'laws of robotics' as an initial thought experiment in the alignment process. These are as follows:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws.

In discussing these three laws, Thomson (2023) comments that 'Asimov missed an essential Fourth Law: A robot must identify itself. We have the right to know if we're interacting with a human or AI' (p.2, original italics). This last point is moot given the fact that certain commentators now speak of AI systems as capable, at least, of simulating human consciousness so as to pass the famous Turing Test in this area (Kleppen, 2023).

Tegmark, however, is rather more critical of the Asimov laws since they can lead to contradictions which are harmful to humans. He suggests the replacement of the three laws with just two in order to codify the autonomy principle for future life forms (2017, p.273):

1. A conscious entity has the freedom to think, learn, communicate, own property and not be harmed or destroyed.
2. A conscious entity has the right to do whatever doesn't conflict with the first law

It may seem strange to think of AI systems as conscious entity but, as is discussed below in relation to non-human persons, this may be a feasible way forward to protect all parties in human/AI collaborative activities.

There is still the problem of ethical alignment to be dealt with and in recent work Bostrom (2023) examines a range of

relevant questions. He is convinced that:

> *To the extent that ethics is a cognitive pursuit, a superintelligence could do it better than human thinkers. This means that questions about ethics, in so far as they have correct answers that can be arrived at by reasoning and weighting up of evidence, could be more accurately answered by a superintelligence than by humans. The same holds for questions of policy and long-term planning; when it comes to understanding which policies would lead to which results, and which means would be most effective in attaining given aims, a superintelligence would outperform humans (p.1)*

Moreover, there is an insistence that:

> *the best way to ensure that a superintelligence will have a beneficial impact on the world is to endow it with philanthropic values. Its top goal should be friendliness. How exactly friendliness should be understood and how it should be implemented, and how the amity should be apportioned between different people and nonhuman creatures is a matter that merits further consideration. I would argue that at least all humans, and probably many other sentient creatures on earth should get a significant share in the superintelligence's beneficence (ibid., p.2)*

This seems fine as far as it goes but it leaves lots more to be done. As Bostrom admits, friendliness is open to many interpretations and there is a need to locate this concept within a framework of universally accepted values. This raises many philosophical questions concerning the origins and nature of human morality though plausible answers are available in the research and literature on this topic. Debate in this sphere tends to be informed principally by arguments based on evolutionary theory about the origins and relative merits of co-operation/competition and egotism/altruism in human behaviour.

Jeremy Griffith (2017) shows how the cruder forms of Social Darwinism which misinterpreted ideas about the struggle for existence were gradually replaced by ideas which demonstrated how moral virtues such as altruism were more beneficial to human society than selfish competition. Dawkins (2017) explains, in what evolutionary psychologists call the environment of evolutionary adaptedness (EEA) it is plausible that – even in a world of fundamentally selfish entities – 'those individuals that co-operate turn out to be surprisingly likely to prosper' (p.58). He goes on to note that:

> *Brains as big as ours...can actively rebel against the dictates of the naturally selected genes that built them. Using language, that other unique gift of the ballooning human brain, we can conspire together to devise political institutions, systems of law and justice, taxation, policing, public welfare, charity, care for the disadvantaged. We can invent our own values. Natural selection gives rise to these only at second remove, by making brains that grow big. From the point of view of the selfish genes our brains raced away with their emergent properties, and my personal value system regards this with a distinctly positive sign (p.61).*

Moral philosophers offer a similar story about the evolution of altruism and co-operation in human society and Harris

(2010) sums up the position succinctly in observing:

> *Clearly, our selfish and selfless interests do not always conflict. In fact, the well-being of others, especially those closest to us, is one of our primary (and, indeed, most selfish) interests. While much remains to be understood about the biology of our moral impulses, kin selection, reciprocal altruism and sexual selection explain how we have evolved to be, not merely atomized selves in thrall to our self-interest, but social selves disposed to serve a common interest with others (pp. 56-57).*

All such arguments would lend further support to co-operative/altruistic thesis, and this case can be reinforced by anthropological and historical/cultural research evidence contained in the work of Christopher Boehm (2012) and Jeremy Lent (2017) on the origins of morality in early societies. Lent described in fine detail how human cultural evolution – and crucially our mores, morals and legal conventions - was irrevocably shaped by the move from hunter-gatherer to agrarian forms of life. Beginning with the Natufian civilisation in the Eastern Mediterranean, settled communities arose in Jordan, Syria and the Lebanon in which tribes started to plant and store grain seeds, build permanent houses, and construct legal conventions concerned with property rights. As Lent summarises such developments:

> *the agrarian worldview transformed the hunter-gatherer's sense of nature as a giving environment into one of a cosmos demanding far more from its human participants, giving birth to a world filled with the existential anxiety that has remained with us ever since (p.104).*

Boehm's (2012) monumental anthropological research work on moral origins traces the evolutionary development of hominids in seeking to explain how genetic and cultural factors combined to shape the emergence of co-operation, generosity and altruism. The central thesis is that:

> *prehistorically humans began to make use of social control so intensively that individuals who were better at inhibiting their own antisocial tendencies, either through fear of punishment or through absorbing and identifying with their group's rules, gained superior fitness. By learning to internalize rules, humankind acquired a conscience... (p.17).*

In commenting upon the move from hunter-gatherer to agrarian settled communities described by Lent (2017), Boehm illustrates graphically how – through the suppression of alpha male behaviour through punishment and social ostracism – evolutionary adaptations to social and economic changes led to a move from a 'wolflike or apelike "might is right", fear-based social order to one also based on internalizing rules and worrying about personal reputations' (p.176).

A particularly interesting and relevant contribution to this ethical debate was made by Oliver Scott Curry (2018) which aligns with the arguments and evidence referred to above. Curry asks the key questions in this sphere:

> *What is morality? And are there any universal moral values? Scholars have debated these questions for millennia. But now, thanks to science, we have the answers. Converging lines of evidence – from game theory, ethology, psychology, and anthropology – suggest that morality is a collection of tools for promoting cooperation (p.40).*

Arguing that morality 'is always and everywhere a cooperative phenomenon', Curry outlines the research by his team which demonstrates that – although there are understandably ethnic, national and cultural variations in ethical codes – our common cultural and biological mechanisms 'provide the motivation for social, cooperative and altruistic behaviour'. The upshot is that seven moral rules are found in codes throughout the world. As Curry explains the finding:

> *as predicted by the theory, these **seven moral rules** – love your family, help your group, return favours, be brave, defer to authority, be fair, and respect others' property – appear to be universal across cultures. My colleagues and I analyzed ethnographic accounts of ethics from 60 societies (comprising over 600,000 words from over 600 sources)2. We found that these seven cooperative behaviours were always considered morally good (ibid., p.41, itlalics added)*

Given this body of research indicating a consensus on seven key moral rules, I suggest that there is sufficient material here to inform the process of aligning AI systems with the beneficent values designed to prevent harm and promote flourishing in both human and non-human persons.

Of course, different nation states – we should consider parallel AI trends in countries such as North Korea, China or Russia – might have different sets of values which are antithetical to the interests of Europe and the USA. In the final analysis, we may have to revert to the default position obtaining in the development of nuclear weapons. Nuclear war has been a constant fear since the first atomic bombs were dropped on Japan in 1945. What has kept the world free of nuclear war during the last 80 years is mutual assured destruction (MAD) based on the fear of retaliation. Jeremy Straub (2019) considers that there may be MAD parallels in the sphere of cyber warfare with deterrence based on realistic fears of mutual attacks. Such a state of affairs may seem far from satisfactory but – given the rapid pace of AI development and deployment – this may end up being the most realistic default position.

AI and Non-Human Personhood

The concept of non-human personhood was initially introduced by Peter Singer (1975/2009) in his groundbreaking work on animal liberation which posited the notion that there are certain conscious entities, particularly animals such as higher primates, who, arguably, have sufficient sentience to warrant a moral status on a par with humans.

Initially, Singer wanted to argue that certain animals should be brought into the moral community on the grounds that they were sentient beings who could suffer. In recent work, Singer has examined the status of AI machines and robots in the light of the wider questions of rights and responsibilities surrounding the human/AI interface. Anticipating contemporary developments, Singer (2023) asks 'if machines can and do become conscious, will we take their feelings into account?'

This is considered a pertinent question since:

> *our treatment of the only non-human sentient beings we have encountered so far – animals – gives no ground for confidence that we would recognize sentient robots not just as items of property, but as beings with moral standing and interests that deserve consideration. (p.382)*

This resonates with the comments of Tegmark and others mentioned earlier about the importance of locating the AI/Human interface within a moral framework. This will be crucial in areas such as education, health, and the workplace in which AI implementation will require collaboration in order to align with best practice in the interests of all parties.

After all, corporations – such as Open AI's multi-billion dollar partner, Microsoft – are, like other suitably accredited collectives, legal persons. In a similar way, the new LLMs and their subsequent iterations can be assigned legal personhood. As Visa Kurki (2019) argues, there is no reason to doubt that 'an AI can function as a legal person, it can be granted legal personhood on somewhat similar grounds as a human collectivity' (p.175). In recent years there have been moves to grant similar rights to the whole of nature, including cloud forests and threatened species (Watts, 2024).

Treating AI tools as non-human persons – in essence including them in the social, political and moral community – may help to reinforce human/AI collaboration with appropriate rights and responsibilities on both sides. Naturally, it might be suggested that – until the trust which comes from the incorporation of appropriate safeguards is established over the years – it makes sense for humans to be the senior, executive partners in any collaborative enterprise.

Upshot

The speed of AI developments over the last few years has understandably created considerable fear and anxiety, and this has spawned an enormous number of apocalyptic conspiracy theories. Taking a pragmatic perspective – steering a middle course between excessive optimism and pessimism about the new technology – it is possible to discern ways in which the potential advantages of using AI tools can be achieved with sufficient safeguards to bring about huge benefits for humankind. If AI applications are treated as non-human persons, it is possible to achieve a collaborative interface in all fields which may help us to solve the major existential problems facing the world in the 21st century.

———

No funding was required for the research and writing of this article.

References

- Aitken, J. (2023). AI could threaten some jobs, but it is more likely to become our personal assistant. The Conversation, May 25, https://theconversation.com/ai-could-threaten-some-jobs-but-it-is-more-likely-to-become-our-personal-

assistant-206297 accessed 18.8.23

- Al-Sibai, N. (2023). Former Google CEO Warns AI Could Endanger Humanity Within Five Years. The Byte. 29.11.23. https://futurism.com/the-byte/eric-schmidt-ai-five-years
- BBC (2023). BT to cut 55,000 jobs with up to a fifth replaced by AI. BBC News, May 18, https://www.bbc.com/news/business-65631168. Accessed 9.10.23
- Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism and Shame*(New York; Basic Books)
- Bostrom, N. (2014).Superintelligence. (Oxford: Oxford University Press)
- Bostrom, N. (2003). *Ethical Issues in Advanced Artificial Intelligence.*https://nickbostrom.com/ethics/ai accessed 7.10.23
- Chen, C. (2023). AI will transform and teaching and learning. Let's get it right. Stanford University AI+ Education Summit, March 9. https://hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right accessed 3.9.23
- Curry, O.S. (2018*). Seven Moral Rules Found All Around the World*, in Price, M., Sloan, M. & Wilson, D.S. (eds)(2018). *This View of Morality: Can an Evolutionary Perspective Reveal a Universal Morality?*,https://evolution-institute.org/new.evolution-institute.org/wp-content/uploads/2018/04/tvol-morality-publication-web2018-7.pdf. pp.40-41
- Dawkins, R. (2017). *Science in the Soul*. (London: Bantam Press)
- Davenport, T. (2019). The potential for artificial intelligence in healthcare. National Library of Medicine, 6(2): 94-98, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/ accessed 22.1.24
- Foucart, R. (2023). Tech giants forced to reveal AI secrets – here's how this could make life better for all.*The Conversation*. April 27,https://theconversation.com/tech-giants-forced-to-reveal-ai-secrets-heres-how-this-could-make-life-better-for-all-204081 (accessed 19.5.23)
- European Parliament (2023). Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/ artificial-intelligence-act- deal-on-comprehensive-rules-for-trustworthy-ai accessed 14.1.24
- Future of Life Institute (2023). Pause Giant AI Experiments – An Open Letterhttps://futureoflife.org/open-letter/pause-giant-ai-experiments/ accessed 21.10.23
- Gibson, K. (2023). Meta sued by states claiming Instagram and Facebook cause harm in children and teens. CBS News. 27/11/23. https://www.cbsnews.com/news/meta-sued-facebook-instagram-children-teens-harm-mental-health/ accessed 17.1.24
- Griffith, J. (2017). *Freedom: The End of the Human Condition*(www. HumanCondition.com)
- Harris, S. (2010) *The Moral Landscape: How Science Can Determine Human Values*(London: Free Press)
- Hassett, J. (2024). The Truth About AI. https://www.youtube.com/watch?v=eFQdWAeZxlE accessed 17.1.24
- Heaven, W.D. (2023). ChatGPT is going to change education, not destroy it. MIT Technology Review, April 6, https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/. accessed 23.1.24
- HSA. (2023). Health and Safety Authority - Irelandhttps://www.hsa.ie/eng/your_industry/health_and_social_care_sector/healthcare_regulators/
- Hyland, T. (2023a). Educational Responses to AI Applications: Problems and Promise. Qeios. August 2023, DOI:

10.32388/08UCQU

- Hyland, T. (2023b). AI Applications and Non-Human Persons. *Philosophy of Education Society of Great Britain,*June 13, https://www.philosophy-of-education.org/ai-applications-and-non-human-persons/
- Kelly, J. (2023). Goldman Sachs Predicts 300 Million Jobs Will Be Lost Or Degraded By Artificial.Intelligence.Forbes.May,31,https://www.forbes.com/sites/jackkelly/ 2023/03/31/goldman-sachs-predicts-300-million-jobs-will-be-lost-or-degraded-by-artificial-intelligence/ accessed 12.9.23
- Khalil, J. (2023). Military, tech experts raise concerns about AI weaponization: 'We have to be very concerned'. The Hill. March 6, https://thehill.com/homenews/nexstar media_wire/4033091-military-tech-experts-raise-concerns-about-ai-weaponization-we-have-to-be-very-concerned/ accessed 15.11.23
- Kleppen, E. (2023). What is the Turing Test? Built In. January 13,https://builtin.com/artificial intelligence/turing-test accessed 25.1.24
- Kurki, V.A.J. (2019*). The Legal Personhood of Artificial Intelligences* (Oxford: Oxford Academic) https://academic.oup.com/book/35026/chapter/298856312?login=false accessed 20.1.24
- Kwo, L. (2021). Contributed: Top 10 Use Cases for AI in Healthcare. MobiHealth News. July 1, https://www.mobihealthnews.com/news/contributed-top-10-use-cases-ai-healthcare accessed 17.1.24
- Lancaster, K. (2023). Why universities should embrace AI essay-writing software. PESGB Blog, February 24. https://www.philosophy-of-education.org/why-universities-should-embrace-ai-essay-writing-software/ accessed 12.1.24
- Lent, J. (2017).*The Patterning Instinct: A Cultural History of Humanity's Search for Meaning*(New York: Prometheus Books)
- Myre, G. (2018). Stanislav Petrov, 'The Man Who Saved The World,' Dies At 77. NPR-The Two Way. https://www.npr.org/sections/thetwo-way/2017/09/18/551792129/stanislav-petrov-the-man-who-saved-the-world-dies-at-77 accessed 13.10.23
- Nicoletti, L. & Bass, D. (2023). Humans Are Biased. Generative AI Is Even Worse. Bloomberg. https://www.bloomberg.com/graphics/2023-generative-ai-bias/ accessed 14.1.24
- Open AI. (2023). Frontier AI regulation: Managing emerging risks to public safety.https://openai.com/research/frontier-ai-regulation
- Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity.* (London: Bloomsbury)
- PubMed (2024). PubMed articles on AI in last 18 months. https://nortonsafe.search.ask.com/web?omnisearch=yes&q=pubmed+articles+on+AI+in+last+18+months accessed 12.1.24
- ResearchGate (2024). ResearchGate articles on AI since November 2022. https://nortonsafe.search.ask.com/web?l=dir&qo=serpSearchTopBox&p2=%5EEQ%5Ezz00ie%5E&ueid=33eb51cb-fb6c-4383-bb49-cca5c38cd927&q=ResearchGate+articles +on+AI +since+November+2022. Accessed 13.1.24
- Rosenblum, A. (2023). Humans have improved at Go since AIs became best in the world. New Scientist. March 13, https://www.newscientist.com/article/2364137-humans-have-improved-at-go-since-ais-became-best-in-the-world/ accessed 20.1.24
- Sanchez-Graells, A. (2023). The UK wants to export its model of AI regulation, but it's doubtful the world will want it. The Conversation. June 7, https://theconversation.com/the-uk-wants-to-export-its-model-of-ai-regulation-but-its-

doubtful-the-world-will-want-it-206956 accessed 15.6.23

- Singer, P. (1975/2009).*Animal.Liberation*(NewYork:HarperPerennial.ModernClassics)
  Singer, P. (2023).*Ethics and the Real World*. (Princeton: Princeton University Press)
- Straub, J. (2019). Mutual assured destruction in information, influence and cyber warfare: Comparing, contrasting and combining relevant scenarios. *Technology in Society.* https://doi.org/10.1016/j.techsoc.2019.101177 accessed 23.1.24
- Taylor, J. & Hern, A. (2023). 'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation. *The Guardian*, 2/5/23, https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning accessed 30.9.23
- Tegmark, M. (2017). *Life 3.0: Being Human in an Age of Artificial Intelligence* (London: Penguin)
- Tegmark, M. (2023). How to Keep AI Under Control. *TED Talk.* https://youtu.be/xUNx_PxNHrY?si=jVwXJXoSnZn0F2xk accessed 3.2.24
- Thomson, J. (2023). 3 rules for robots from Isaac Asimov — and one crucial rule he missed.*Big Think*, 9/3/23. https://bigthink.com/the-future/3-rules-for-robots-isaac-asimov-one-rule-he-missed/ accessed 28.1.24
- Watts, J. (2024). Could 2024 be the year nature rights enter the political mainstream?*The Guardian*, 1/1/24
- White House. (2023). Blueprint for an AI Bill of Rights.https://www.whitehouse.gov/ostp/ai-bill-of-rights/ accessed 11.10.23
- Wiggers, K. (2024**).** OpenAI claims New York Times copyright lawsuit is without merit. *TechCrunch*. 8.1.24. https://techcrunch.com/2024/01/08/openai-claims-ny-times-copyright-lawsuit-is-without-merit/?guccounter=1 accessed 25.1.24
- World Economic Forum. (2023). *These are the jobs most likely to be lost – and created – because of AI*May 4, https://www.weforum.org/agenda/2023/05/jobs-lost-created-ai-gpt/ accessed 12.11.23
- Yang, M. (2023). Scientists use AI to discover new antibiotic to treat deadly superbug.*The Guardian*. May 25, https://www.theguardian.com/technology/2023/may/25/artificial-intelligence-antibiotic-deadly-superbug-hospital accessed 28.11.23