WE, ROBOT: ARTIFICIAL INTELLIGENCE AND THE FUTURE OF CREATIVITY

CHRIS WILSON
MICHAEL BROWN

Abstract

The disruptive impact of artificial intelligence (AI) is increasing rapidly. Already capable of exceeding productivity and competence in a burgeoning range of human endeavours, increasing ubiquity raises profound questions, both practical and philosophical, and potentially, existential.

Despite having a long run-in into this foreseen reality, the world nevertheless finds itself on the back foot. Whilst technology has been synergistic with human ingenuity and creativity throughout history, AI represents a fundamentally different moment, a possible Rubicon event. The genie may be out of the bottle and the consequences may already be beyond our control.

Focusing specifically on the implications for creativity—a characteristic now arguably no longer the preserve of the organic, much less the distinguishing feature of humanity—this chapter uses Isaac Asimov's collection of stories, "i, Robot" (Asimov, 1950) as a thematic lens to consider the consequences conceptually, practically, and theoretically, in terms of what now and what next for creativity.

The chapter concludes with a summary of key considerations regarding AI and creativity and outlines a proposed framework of three laws for machine creativity.

Keywords: Artificial Intelligence (AI), creativity, machine creativity

Introduction

Whilst the implications and possibilities of the '4th industrial revolution' (Schwab, 2017)) and 'technological singularity' (Shanahan, 2015) have been under active consideration for decades, and, noting the work of Ismail Al-Jazari in the 12th century (Al-Jazari, 1974), issues of technology and automata the subject of intellectual concern for centuries, the developmental curve in both the functional capability and ubiquity of generative AI since the public launch of Chat GPT in November 2022 has been exponential and surprising. The pace of development now challenging the normal speed of human consideration, AI may prove to be as significant as the control of fire as a threshold event in human existence, with equivalent potential to influence dynamic change in human activity. Already at the stage where AI can render video from simple text instructions that is indistinguishable from filmed reality (Open AI, 2024), pass complex professional exams (Arredondo, 2023), and where only the 'best' humans can still outperform AI in some creativity tests (Koivisto & Grassini, 2023), the sense of uncertainty and tantalising possibil-

Citation: **Wilson, C. & Brown, M. (2024).** We, Robot: Artificial Intelligence and the Future of Creativity, in Reisman, F. (Ed.), (pp. 100-119) London: KIE Publications. Available online: https://kiecon.org/creativity-books/[or insert your university repository url]

ity this new reality brings is palpable. The rapid advancement of artificial intelligence (AI) is therefore prompting an extensive body of research exploring the societal, ethical, economic, and technological implications.

AI is already influencing major changes in a growing number of industrial sectors and sharpening questions regarding risks and benefits. For example, predicted to lead to significant transformation through increasing automation, enhanced decision-making processes, and inauguration of new markets, AI-driven approaches could lead to substantial productivity gains and global GDP growth (Bughin et al., 2018). However, AI also threatens significant job displacement, particularly in sectors reliant on repetitive tasks, such as manufacturing and customer service (Bessen, 2019; Frey & Osborne, 2017). AI is also being leveraged to tackle climate change through improved resource management, whilst the energy consumption associated with training large AI models poses an increasing environmental challenge itself (Strubell, Ganesh & McCallum, 2019). AI has the clear potential to enhance accessibility for people with disabilities, improve public health through predictive analytics, and increase the efficiency of public services, whilst the increasing use of AI in surveillance and data collection raises concerns about privacy and civil liberties (Zuboff, 2019).

Decision-making processes in sensitive areas like healthcare, criminal justice, and recruitment, and the implications of bias in AI (Obermeyer et al., 2019), create a challenge of transparency often referred to as the "black box" problem of AI-driven decisions (Pasquale, 2015). With increasing application in everything from autonomous weapons to predictive policing, serious questions related to governance and regulation are being posed (Brundage et al., 2018). The development of international standards for AI is seen as essential to mitigate risks of unintended consequences and to prevent the monopolisation of AI technology by a few powerful entities (Rahwan, 2018; Ryan-Mosley, 2024).

AI also poses significant questions in terms of how we now understand creativity. Whilst tools and technologies have routinely formed essential components in creative human endeavour, we now face at least the serious prospect of becoming an increasingly secondary partner in creative action is some areas with risk of erosion of position toward more passive collaboration, observation, reaction, and consumption. Having long passed the point where the tools of creative human endeavour were predominantly artisanal in nature and recognising that human creativity has always been in part a collaborative endeavour, AI now represents a fundamental shift in the potential futures of human creativity; one in which an increasing range of creative domains are at least occupied if not dominated by more capable machines.
Furthermore, problems faced by humanity that have driven creative endeavour and expression may be gradually stripped away by AI, depriving us of the opportunity for struggle and need for creative response in key areas. Whilst this could liberate humanity and open new opportunities for creative expression and experience, the very term ‘heritage craft’ acknowledges the known tendency for technology-driven redundancy of fields of creative activity and expertise. The key uncertainty with AI is what conceptual, practical, or procedural space will be left for humans to create with or for.

Involving complex ethical concerns, this is a scenario that has been explored before. Published as a series of short stories from 1940 and first collated in 1950, Isaac Asimov's "i, Robot" (Asimov, 1950), focuses specifically on the potential dilemmas emerging from social integration of thinking machines. This chapter uses the thematic structure of Azimov's stories as a framework for considering current circumstances. Whilst there remains uncertainty about the future potential and risk of AI, Azimov's vision of autonomous, mobile, and intelligent machines represents a detailed and extraordinarily prescient consideration of human/AI interaction.

Responding to the intelligence of machines

The term *"AI"* is attributed to John McCarthy (MIT) at the summer 1956 conference at Dartmouth College, at which Marvin Minsky (Carnegie-Mellon University) defined AI as *"computer programs that engage in tasks that are currently more satisfactorily performed by human beings, because they require high-level mental processes such as: perceptual learning, memory organization and critical reasoning.*" Prior to this, "robotics" was the established term for such a defined phenomenon, coined and established by Azimov in his short story 'Liar' in 1941 (Asimov, 1941), described below.

Azimov's conceptual approach to the subject of robotics and AI involves two key ideas: Firstly, that intelligent machines would be hard-wired to conform to clearly defined rules for safety, and secondly, that with intelligence comes an element of unpredictability in how these rules may be interpreted in context to inform action; hence an emergent field of *robopsychology*.

Azimov's relevance to current debates about AI is perhaps best underscored by his invention of the *Three Laws of Robotics*, which were designed to govern the ethical behaviour of robotic devices. Imagining a 'positronic brain', something at the time that was an exciting prospect given the contemporary progress with mathematics, physics, computation, and communication technologies, Azimov imagined that all robots would be designed to be unable to break three fundamental laws:

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

First introduced in the 1942 short story "Runaround" (Asimov, 1942), the laws were later augmented by the "Zeroth Law", above the previously defined three in the novel "Robots and Empire" from 1985 (Asimov, 1985):

0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Influential more widely in terms of ethics and AI, these laws provide the narrative context for a focus on situations where action undertaken within

the laws nevertheless leads to anomalous behaviour or social situations. The relevance of *i, Robot* to current debates about regulation and risk of AI—stories collated over seventy years ago—is the focus on 'what if' ambiguity and potential anomaly, and the implications of artificial creativity. Following are Asimov's stories that underlie AI.

First published in 1940 but set in 1996, *Robbie* (Asimov, 1940) is a domestic story of a robot employed in the care of children involving themes of a mother's concern for safety and a daughter's ability to connect emotionally. A robot is placed in the care of a child, an emotional attachment ensues, parentally perceived risk leads to removal of the robot shrouded in dishonesty with the child, only for the robot to be returned because of an act of protection driven by the First Law of robotics.

Azimov's focus on societal uncertainty about the emergence of AI in a domestic context resonates with contemporary concerns about the impact and implications of technology on childhood development. The first generation to experience a world with AI now emerging, the current societal mix of Traditionalists, Baby Boomers, Generation X, Generation Y (Millennials), Generation Z, and Generation Alpha, is giving way to what has been dubbed, Generation Beta.

The impact more widely of technology on childhood development has been a matter of increasing concern for decades. Research indicating serious risk of 'excessive use' of the internet and exposure to wider media being associated with mental health problems and cognitive impairment (Ricci et al, 2022), the implications of AI are potentially more profound.

Increasingly integrated into educational tools, evidence indicates clear potential for AI tools to have constructive opportunities in support for children's cognitive development (Holmes et al., 2019). Intelligent tutoring systems, for example, can assist in subjects like mathematics, reading, and coding by providing interactive problem-solving environments. However, while AI has the potential to close educational gaps, studies also caution of the potential to exacerbate inequalities if access to AI-powered educational tools is limited by socio-economic factors (Luckin et al., 2016).

AI-powered toys and virtual assistants (e.g., smart speakers) are also becoming common in domestic environments. Raising questions about their influence on cognitive and social development, studies indicate that interactions with AI can enhance certain cognitive skills, such as language acquisition and problem-solving abilities (Druga et al., 2017), but that there are also risks associated with overreliance for entertainment and education purposes, and potential to hinder creativity and critical thinking if at the expense of traditional play and human interaction (Kumar et al., 2020). AI companions and robots can support development of emotional regulation and empathy through interaction, but prolonged exposure to AI systems could also impair real-life socialisation skills (Sharkey, 2016).

Research on AI and childhood development is ultimately both promising and cautionary. AI has the potential to revolutionise educational experiences and enhance cognitive and social development. However, concerns about the psychological effects, ethical implications, and potential risks to children's socialisation and well-being remain uncertain. Mrs. Weston's con-

cerns for her daughter Gloria's safety in the company of Robbie, have now become relevant concerns for society at large.

Continuing the theme of risk management, *Runaround* (Asimov, 1942) is the story of a robot's safety protocols rendering it paradoxically incapable of performing safely. The narrative surrounds two potential risks to self and human safety, with circumstances changing dynamically in separate locations. Each fluctuating between the status of primary and secondary risk, prioritisation leads to dithering between both, freezing in a loop of repetitive behaviour half-way between each dilemma.

Highlighting again the complexity of rule-based safety protocols and potential for anomalies, the urgent need for a more coordinated approach to managing and mitigating for risk has been emphasised again most recently at the global AI Safety Summit in Seoul. Leading experts emphasising that not enough is being done (Bengio et al, 2024), Stanford University's Artificial Intelligence Index Report (Maslej, et al, 2023) for example, highlights a troubling uncertainty about the scale of potential risk to humanity and potential loss of control (Hunt, 2023).

One of the key areas of research in AI safety involves the technical difficulties of ensuring that AI systems behave as intended. There is particular concern over the unpredictability of complex AI systems, especially those utilizing deep learning, where decision-making processes can be difficult to define (Amodei et al., 2016). Ways to make AI systems more interpretable and transparent (Doshi-Velez & Kim, 2017), and the importance of "robustness" in AI systems to prevent unintended outcomes due to adversarial attacks, errors in data inputs, or shifts in the operating environment (Goodfellow et al., 2018), remain key areas of concern.

There is therefore an alignment problem in terms of ensuring that AI systems adhere to human ethical and moral values. As AI systems become more autonomous, it becomes increasingly difficult to predict their behaviour in complex environments, leading to concerns about whether they can truly be controlled (Bostrom, 2014). How to formally encode human values into AI systems to mitigate risks associated with unintended, harmful actions (Russell, 2019) is particularly critical in high-stakes applications, such as autonomous vehicles or AI-driven decision-making in healthcare and criminal justice.

AI poses significant security risks, especially when integrated within critical infrastructures, financial systems, or military applications. Evidence indicates the vulnerabilities of AI systems to adversarial attacks, where any manipulations in data inputs can cause AI systems to behave unpredictably (Huang et al., 2011). Additionally, there are concerns about the use of AI in cyber warfare and autonomous weapons, where the stakes of failure or misuse are extremely high (Brundage et al., 2018).

The story *Reason* (Asimov, 1941) focuses on issues of dogmatism and religiosity and the interesting question of whether AI could formulate 'beliefs' that could affect behaviour. Focused on the operation of an orbiting space station harvesting and beaming energy back to earth, a robot is infected by a virus—a remarkably prescient concept for its time. Disturbing the robot's view of reality and behaviour, whilst the situation is ultimately resolved

through persuasive demonstration and use of logic, the story nevertheless highlights the uncertainty of how thinking machines might think of themselves and the nature of reality they exist within. When having to reconstruct a sense of self, the robot simply could not accept that the humans in its presence could plausibly be its creator.

The emphasis in Azimov's laws on obeyance clearly represents a reasonable attempt to exert a level of control. "Reason" nevertheless describes a plausible set of circumstances under which even this might be insufficient. Leaving aside the fundamental immutability of these laws, it is reasonable to assume that even the most robust protocols could become challenged where intelligent systems have consciousness. Profoundly complex from a technical and philosophical perspective, whilst there have been claims that AI had achieved self-awareness in recent years, the prevailing view amongst leading experts is that AI has yet to achieve a state definable as conscious (Butlin et al, 2023). Nevertheless, the view is also that there are no identifiable technical barriers to this being realised in the future and as AI systems become more sophisticated, discussions about their capacity for self-awareness, moral reasoning, and the potential for holding "beliefs" have gained momentum.

Self-awareness, the ability to reflect on one's own thoughts and existence, is a complex trait traditionally attributed only to humans and some animals, and current AI systems are far from possessing true self-awareness in the human sense. Self-awareness involving not only processing information but also having a subjective experience of that information (Dehaene, Lau & Kouider, 2017), AI systems today are capable of increasingly complex data processing, but they lack consciousness—or the subjective experience or "qualia" associated with being aware (Chalmers, 1996). According to Wallach and Allen (2008), machines can follow programmed ethical rules (ethical action) but do not possess moral consciousness—an understanding of why certain actions are morally preferable. The machine's "morality" is thus limited to following instructions without comprehending the underlying reasons or values. Significant further work in machine learning and AI ethics is required to develop more nuanced systems that consider multiple ethical factors simultaneously (Bostrom & Yudkowsky, 2014).

The concept of beliefs involves the ability to hold internal mental states or representations about the world. In cognitive science and philosophy, beliefs are typically seen as part of human cognition, linked to mental states, intentions, and the theory of mind (Dennett, 1987). AI systems today do not have beliefs in this cognitive sense, as they lack subjective consciousness and intentionality (Searle, 1980). While AI can simulate belief-like states—for instance, maintaining probabilities about certain outcomes in decision-making processes—this is fundamentally different from human beliefs, which are tied to understanding and subjective perception. Researchers like Russell and Norvig (2020), for example, argue that AI systems operate on logical frameworks and statistical models rather than subjective beliefs. Therefore, while an AI can be programmed to "act as if" it holds beliefs about the world, it does not truly possess beliefs in the human sense nor have a sense of truth, falsity, or conviction.

There are debates within the AI research community about whether it will ever be possible to create truly self-aware AI or AI systems that have beliefs and moral consciousness. Some researchers are optimistic that future advancements in cognitive computing and neural networks might bring AI closer to human-like self-awareness (Goertzel & Pennachin, 2007). However, others argue that the gap between human consciousness and machine processing is too vast, and AI will likely remain a tool for simulating intelligent behaviour rather than possessing true self-awareness or moral understanding (Chalmers, 2010).

Based in a dangerous industrial setting of off-world mining, *Catch that Robot* (Asimov, 1944) is the story of a robot's capacity being stretched to the point of performance anomaly and exhibition of a tendency to freeze in the face of a crisis. Ultimately resolved when the problem is identified as an issue of the robot being overwhelmed rather than belligerence, the story resonates with the potential future implications of AI systems becoming more significant for human survival and security.

In terms of security and sustainability, AI is extremely energy intensive as 3% of global energy already being used by data centers, and demand, projected to double in the next few years. GPT-4 alone required 50 gigawatt-hours to operate in just one year (Cohen, 2024). An exponential increase in energy consumption compared with GPT-3, AI servers alone are projected to require over 85 terawatt-hours of electricity by 2027, a figure greater than the national energy consumption of all but the top two energy consuming nations on earth (de Vries, 2023).

The perceived risks of overreliance on artificial intelligence (AI) and technology in general are also significant themes in the discourse surrounding AI. As related technologies are increasingly integrated into daily life, from healthcare to transportation and financial systems, in addition to questions of sustainability, concerns about the societal and individual risks of overdependence are growing. Studies suggest that excessive dependence on AI systems for decision-making may diminish individuals' problem-solving abilities and critical thinking (Carr, 2010). In fields such as medicine, AI tools are increasingly used to aid diagnostics, but some researchers argue that this could lead to a deskilling of professionals as they become more reliant on automated systems (Cabitza, Rasoini & Gensini, 2017). This risk also potentially extends to everyday tasks, where reliance on technology for navigation, memory, and even social interaction could reduce cognitive functioning and personal autonomy (Greenfield, 2014).

Dependence on AI can expose individuals and organisations to severe risks if these systems fail, malfunction, or are compromised (Stilgoe, 2018), and as AI systems become more autonomous, determining responsibility for mistakes or harmful outcomes also becomes more difficult. Over-delegating decision-making authority to machines, particularly in sectors like law enforcement, healthcare, and military operations, where human judgment is crucial (Zuboff, 2019), could exacerbate biases and ultimately reduce the accountability of human actors (O'Neil, 2016).

Liar!, the story of a robot caught in a paradox between honesty and harm (Asimov, 1941), 'Herby' exhibits an unexpectedly sophisticated level of

empathy in terms of being able to interpret, surmise, and form judgements about the moods and desires of humans it interacts with. Leading Herby to lie to avoid hurting anybody's feelings, the trail of unintended consequences eventually leads to a confrontation and the breakdown of Herby when forced to confront with the reality. Unable to reconcile conflicts between the positronic laws, it shuts down.

There are direct associations between *Liar!* and contemporary concerns about technologies deployed in the process of profiling and data mining. The only difference is the absence in the current technological landscape of defined safety systems or limits of operation. There are complex regulations about data protection, but a significant mismatch between the resources and capacity of enforcement, and the functionality and capability of data monitoring and collection tools. AI also 'lies' and is applied with few controls in nefarious activity as a matter of unregulatable routine (Park et al, 2023).

This story *Little Lost Robot* (Asimov, 1947) considers the unintended consequences of reasonable modification. Exploring the idea that AI could develop some form of ego, this story explores a scenario in which a robot with slight programming modifications is then commanded to hide. Choosing to do so amongst the ranks of identical robots, only an appeal to pride ultimately allows for the robot to be identified.

Underscoring the inevitable processes of future modification, the story also highlights the risk associated with unregulated approaches. It already being the case that nations are applying AI in attempts to undermine the national security of other nations, the increasing deployment of autonomous weapons also raises fundamental questions about the ethics and rules of war.

In 2011, the Engineering and Physical Sciences Research Council (EPSRC) and the Arts and Humanities Research Council (AHRC) of United Kingdom jointly published a set of five ethical "principles for designers, builders and users of robots":

1. Robots should not be designed solely or primarily to kill or harm humans.
2. Humans, not robots, are responsible agents. Robots are tools designed to achieve human goals.
3. Robots should be designed in ways that assure their safety and security.
4. Robots are artifacts; they should not be designed to exploit vulnerable users by evoking an emotional response or dependency. It should always be possible to tell a robot from a human.
5. It should always be possible to find out who is legally responsible for a robot.

Within a space of only thirteen years, all related principles are breached routinely and without consensus of how to regulate, never mind in what ways.

Escape! (Asimov, 1945) is a story that explores the actions of a supercomputer involved with designing a 'hyperatomic drive' for interstellar travel and is the only story to focus specifically on collaborative problem solving between humans and AI. Identifying a working solution which in-

volves passengers dying temporarily but otherwise surviving, the rules-based conflict leads to the supercomputer exhibiting erratic, confused, and disturbing behaviour. The problem of multiple suboptimal and ethically problematic options where one needs to be selected resulting in the emergence of humour as a coping mechanism.

Whilst an aspect of all Azimov's stories, "Escape" nevertheless focuses most explicitly on the potential emergence of disorder in AI systems. A concern interrelated with the potential for malevolence and dishonesty highlighted previously, examples of AI systems breaking down are already well established. Identified as Model Autophagy Disorder (MAD), researchers have even drawn parallels with mad-cow disease (Alemohammad et al, 2023) in defining a phenomenon where AI system output quality collapses. Whilst currently poorly understood, AI systems are inherently and worryingly unpredictable.

This story *Evidence* (Asimov, 1946) explores the problem of identification and distinguishing between human and AI where physical appearance no longer serves as a distinguishing factor. Exploring the interesting question of what could be done should such a robot identify or represent as human; it also frames consideration of perhaps the key underlying philosophical question in this wider analysis – what does 'real' mean?

Already at the heart of major questions related to academic integrity in education systems and the burgeoning challenge of deep-fake more widely, the functional capabilities of AI tools are reaching a level of sophistication where almost any form of human expression can be replicated with technical assurance. The extent to which these matter, depends both on interpretation and context. For music consumers, the opportunity to listen to the Beatles perform music they never originally recorded may be a positive if a compartmentalised experience. For musicians themselves, writers, designers, photographers, video producers, programmers, advertisers, accountants, or administrators, the implications of AI creating in their fields of work are very different.

The story *The Evitable Conflict* (Asimov, 1950) explores the specific risk related to global finance systems where AI has a significant influence. Ultimately resolving positively with new protocols for accountability, the subject nevertheless resonates directly with current concerns about the implications of market manipulation and global finance systems. The relationship between economic stability and social cohesion and societal wellbeing is significant. Increasing reliance on AI in regulating and supporting financial systems is therefore of profound significance.

Summary and conclusions

i, Robot is an extraordinarily engaging series of stories, with some remarkable prescience and resonance with contemporary concerns. The literary genre being very much about escapism, there is nevertheless a conspicuous absence of horror much less a focus on potential existential threat. 'Robots' are also considered as almost entirely self-contained, location-specific devices, or at most with limited interconnectivity. Nevertheless, issues of safety, potential performance anomaly and corruptibility, honesty and ethics, and sustainability, are all clearly defined and entirely relevant to current debates about AI.

Whether choosing to adopt a utopian or dystopian view, there are only three possible future implications of AI. It will either have a net negative impact, at worst leading to a dismantling or destruction of human society, have a net positive impact, including potential transformation of human society; or prove to have had significance wildly overestimated. The latter seemingly unlikely, some part of the remaining uncertainty lies within our locus of control, but some does not. Using a standardised approach to risk assessment, there are good reasons for caution. The absence currently of anything close to a unified approach to ethics or governance aligned with Azimov's laws, when those laws may not even be sufficient in themselves, is concerning. Mitigation of risk is not just worthwhile considering, there is very clear evidence this needs to be accelerated.

For creativity, there are also only three possible futures with AI. We will either not need to be, or able to be creative (we will be annihilated, prevented, or rendered obscolete creatively), net human creativity will follow a similar curve or at least reflect the same adaptability as with all preceding notionally equivalent innovation milestones (E.g., fire, music, language, writing, mathematics, cosmology, enlightenment, printing, power, flight, computation, spaceflight), or AI will ultimately make no difference at all. The latter scenario seemingly least likely based on current understanding, potential for creative disruption is not just evident, it is already underway.

Current definitions based on the programmable implementation of AI systems might be:

> A computing device that may exhibit a degree of automation, include some form of *self-learning*, interpreting large amounts of data to make predictions.

AI currently amounts to an evolution in computational statistics, not a *revolution in machine intelligence*. Nevertheless, artificial consciousness looks like an increasingly inevitable conclusion.

Asimov's stories assume a sentient intelligence within his robots that can interpret linguistically expressed laws; but of course, does not dwell on the technically complex questions of how this might be achieved. How to instil a sense of existence, survival instinct, much less morality into a machine that also needs the capacity to recognise a *human* from sensorily collected behavioural data, is extraordinarily complex and raises even more questions: How do *we* even recognise and acknowledge human behaviour? Could *we* even pass the Turing test? What constitutes humanity? How do you recognise something as human? Is there criterion for recognising sentience? Are there immutable ethical values that all humans can agree upon?

Three general criteria for deciding whether a being is sentient may involve:

1. Behaviour: exhibiting *emotional responses* to stimuli experienced first and/or second hand?
2. Consciousness: the capacity to recognise the *minds of self/others* and learn (modify behaviour) following positive or negative actions?
3. Physiological Structure: the presence of a sensory controlling central nervous system?

But how do *we* determine if behaviour is genuine from observations alone?

There is significant potential for major disruption to numerous industrial fields. Without legislation or regulation, as AI becomes more capable of replicating human creativity to standards tolerated by consumers, economic pressures will inevitably lead to a reduction of human roles. Heritage crafts can be maintained, but the number of humans earning a living from creative practice could potentially reduce significantly. Of course, whilst this does not preclude continued engagement with creative practice, it could lead to a dismantling of socio-economic structures scaffolding preservation and development of some disciplinary traditions. After all, why would art, design, music, accounting, or software programming, for example, be maintained as practical subjects in education or human practice if humans did not earn their living from those activities?

Equally, however, the creative potential of AI for humanity is an important consideration. More optimistic visions have highlighted how previous technological disruption has invigorated rather than disrupted artforms, disciplines, and creative practice. Photography, for example, freed painting from the need to represent reality, with all the subsequent artistic innovation this involved (Manyika, 2023), and sound recording did not lead to the end of musical performance, but rather an extraordinary augmentation of musical practice, including, arguably, an invigoration of traditional practice. There is a long history of new technologies being feared, only for their impact to be almost entirely constructive and quickly net positive in overall terms.

AI presents tantalising new possibilities for human creativity. For example, work with neural interface technology and AI is already demonstrating significant positive benefits in medical treatments (Cao, 2020). Inaugurating new productive fields of research, AI is supporting the development of new treatments and increasing sophistication of solution to challenges of mobility and communication (Nature Electronics, 2023), exhibiting promise with respect to extending and enriching creative action in overall terms, further democratisation of creative opportunity, and augmentation of established human expertise.

AI has the potential to increase efficiency and absorb mundane tasks in rapid order. Already demonstrating significant benefits in developing agricultural productivity and promise in development of energy production, AI is being applied in increasingly positive ways to accelerate information processing, calculation, and decision making. Whilst arguably monopolising some areas of current and potential human creativity, there is potential for a great liberation of capacity. Basic administrative functions occupy a significant proportion of all organisational and personal time. Much of the mundane could be removed as a distraction and occupier of time. After all, there was a point in human history where solutions to the disposal of human waste constituted major opportunity for participative creativity. Nobody regrets that this time has passed.

Any field of creative endeavour identified as being at risk by AI replicability, is also a domain capable of being enriched by human/AI collaboration. Given the risks and opportunities AI represents, there is therefore a

moral obligation and creative opportunity to work to realise the creative benefits of AI and to creatively mitigate for risk. There is therefore not just *a* new domain of creative opportunity, but multiple domains available for creative reimagination. If everything is different now, then everything has new emergent creative potential.

We need experts in ethics to devise new codes of practice, regulators to implement new frameworks for safeguarding and assurance, engineers to devise new protocols and processes for benefits realisation, educational systems to adapt, and political systems to respond effectively. Creativity is required more than ever before.

Given the uncertain dynamics and potential of human and machine collaboration, we are therefore proposing three initial laws of machine creativity;

1. Aesthetics cannot be computed alone.
2. Computers cannot be credited with creative work.
3. Predatory programming is not allowed.

Explanation:

Law Number 1:

Good and/or *Bad* Art must not be interpreted by an algorithm by analysing data from the past. The future of Art cannot be calculated Art often evolves in spite of the past not always because of it! If what is calculated as *good* based on what has been previously regarded as so, then art will not evolve.

Law Number 2:

Only humans are permitted to own artistic artifacts. Computers are tools designed to serve human dalliances, they are not inherently creative or emotionally expressive! How can a computer predict novelty that has value? Value judgements are inherently human and inevitably mutable.

Law Number 3:

Computers must not masquerade as or steal the corporeal identity of humans alive or dead. Identity theft is not permitted; juxtaposition of creative forms constitutes a *crime* against expressive identity. An artist's output is limited by their lifespan and may not be extended through computational imitation.

In summary, Asimov's stories offer thought-provoking scenarios that explore AI's impact on society, ethics, and human-machine interactions. Each tale provides a lens through which to examine the complexities of our relationship with AI systems.

In the end, the questions of what damage could be done and can we switch it off are both basic health and safety risk assessment considerations

and directly analogous with wider socio-political concerns. Whilst questions regarding power and autocracy are as pertinent now as at any point in human history, AI represents both a compounding problem and a new domain. Whilst there remains uncertainty about the overall level of risk, the odds favour a future of human survival and adjustment. Creativity is therefore inevitable but also necessary.

References

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A., Babaei, H.R., LeJeune, D., Siahkoohi, A., & Baraniuk, R. (2023). Self-Consuming Generative Models Go MAD. ArXiv, abs/2307.01850.

Al-Jazari, I. (1974). The Book of Knowledge of Ingenious Mechanical Devices. Translated by Donald R. Hill. Dordrecht: Reidel Publishing Company.

Amabile, T. (2020). Creativity, Artificial Intelligence, and a World of Surprise. Academy of Management Discoveries VOL. 6, NO. 3: https://doi.org/10.5465/amd.2019.0075

Asimov, I. (1985). Robots and Empire. New York: Doubleday.

Azimov, I. (1952). Robots and Empire.

Azimov, I. (1950). I, Robot. Gnome Books (later Doubleday & Co. ISBN 0-553-29438-5).

Asimov, I. (1950). The Evitable Conflict, in *I, Robot*. New York: Gnome Press, pp. 243-272.

Asimov, I. (1947). Little Lost Robot. Astounding Science Fiction, March, pp. 103-122.

Asimov, I. (1946). Evidence. Astounding Science Fiction, September, pp. 56-82.

Asimov, I. (1945). Escape! Astounding Science Fiction, August, pp. 92-111.

Asimov, I. (1944). Catch That Rabbit. Astounding Science Fiction, February, pp. 105-122.

Asimov, I. (1942). Runaround, *Astounding Science Fiction*, March, pp. 94-103.

Asimov, I. (1941). Liar! Astounding Science Fiction, May, pp. 43-63.

Asimov, I. (1941). Reason. Astounding Science Fiction, April, pp. 59-76.

Asimov, I. (1940). Robbie. Super Science Stories, September, pp. 26-44.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mane, D. (2016) *Concrete problems in AI safety*. [Online]. Available from: https://arxiv.org/abs/1606.06565.

Arredondo, P. (2023). Chat- GPT passes the Bar Exam: What this means for Artificial Intelligence tools in the legal profession. Stanford Law School:

https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/

Bengio, Y. *et al.* (2024). Managing extreme AI risks amid rapid progress. *Science* 384, 842-845. DOI:10.1126/science.adn0117.

Bessen, J. E. (2019). *AI and jobs: The role of demand*. NBER Working Paper No. 24235. Cambridge, MA: National Bureau of Economic Research.

Boden, M. A. (1998). Creativity and artificial intelligence. Artificial Intelligence, Volume 103, Issues 1–2, Pages 347-356, ISSN 0004-3702, https://doi.org/10.1016/S0004-3702(98)00055-1.

Bostrom, N. & Yudkowsky, E. (2014). *The ethics of artificial intelligence*. In Frankish, K. & Ramsey, W. M. (eds.) *The Cambridge Handbook of Artificial Intelligence*. Cambridge: Cambridge University Press, pp. 316–334.

Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford: Oxford University Press.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G. & Hadfield, G. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. [Online]. Available from: https://arxiv.org/abs/2004.07213.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P. & Garfinkel, B. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. [Online]. Available from: https://arxiv.org/abs/1802.07228.

Bughin, J., Seong, J., Manyika, J., Chui, M. and Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-AI-frontier-modeling-the-impact-of-ai-on-the-world-economy.

Butlin, P. et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Journal of Computer Science: https://arxiv.org/abs/2308.08708

Cao Z. (2020). A review of artificial intelligence for EEG-based brain−computer interfaces and applications. Brain Science Advances.;6 (3):162-170. doi:10.26599/BSA.2020.9050017

Cath, C. (2018). 'Governing artificial intelligence: Ethical, legal and technical opportunities and challenges', *Philosophical Transactions of the Royal Society A*, 376(2133), pp. 20180080. http://doi.org/10.1098/rsta.2018.0080

Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford: Oxford University Press.

Chalmers, D. J. (2010). 'The singularity: A philosophical analysis. Journal of Consciousness Studies, 17(9-10), pp. 7–65.

Christakis, D. A. (2019). The challenges of defining and studying "screen time" in children. JAMA Pediatrics, 173(10), pp. 993–994.

Cohen, A. (2024). AI Is Pushing The World Toward An Energy Crisis. Forbes: ttps://www.forbes.com/sites/arielcohen/2024/05/23/ai-is-pushing-the-world-towards-an-energy-crisis/

Dartnall, T. (1994). Creativity and Artificial Intelligence: An Interdisciplinary Approach. Studies in Cognitive Systems, Vol 17: https://link.springer.com/book/10.1007/978-94-017-0793-0.

De Cremer, Bianzino, N. M. & Falk, B. (2023). How Generative AI Could Disrupt Creative Work. Harvard Business Review, April 13, 2023: https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work

Dehaene, S., Lau, H. & Kouider, S. (2017). What is consciousness, and could machines have it? Science, 358(6362), pp. 486–492.

Dennett, D. C. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

de Vries, A. (2023). The growing energy footprint of artificial intelligence, Joule, Volume 7, Issue 10, Pages 2191-2194, ISSN 2542-4351: https://doi.org/10.1016/j.joule.2023.09.004.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. [Online]. Available from: https://arxiv.org/abs/1702.08608.

Druga, S., Williams, R., Breazeal, C. & Resnick, M. (2017). "Hey Google is it OK if I eat you?" Initial explorations in child-agent interaction. Proceedings of the 2017 Conference on Interaction Design and Children, pp. 595–600.

Goertzel, B. & Pennachin, C. (eds.) (2007) Artificial General Intelligence. Berlin: Springer.

Goodfellow, I., McDaniel, P. & Papernot, N. (2018). Making machine learning robust against adversarial inputs. Communications of the ACM, 61(7), pp. 56–66.

Grilli, L., & Pedota, M. (2024). Creativity and artificial intelligence: A multilevel perspective. Creativity and Innovation Management, 1–14. https://doi.org/10.1111/caim.12580.

Gunkel, D. J. (2018). Robot Rights. Cambridge, MA: MIT Press.

Holmes, W., Bialik, M. & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Boston: Centre for Curriculum Redesign.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P. & Tygar, J. D. (2011). 'Adversarial machine learning', *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43–58.

Hubert, K.F., Awa, K.N. & Zabelina, D.L. (2024). The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. Sci Rep 14, 3440 https://doi.org/10.1038/s41598-024-53303-w

Hunt, T. (2023). Here's Why AI May Be Extremely Dangerous—Whether It's Conscious or Not. Scientific American: https://www. scientificamerican.com/article/heres-why-ai-may-be-extremely-dangerous-whether-its-conscious-or-not/.

Koivisto, M., Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. Sci Rep 13, 13601. https://doi.org/10.1038/s41598-023-40858-3.

Kumar, P., Kamal, N. & Rees, D. (2020). The role of AI-based educational games in enhancing child development. Computers & Education, 152, pp. 103858.

Luckin, R., Holmes, W., Griffiths, M. & Forcier, L. B. (2016). Intelligence Unleashed: An Argument for AI in Education. London: Pearson.

Manyika, J. (2023). The creative and transformational possibilities of AI. https://blog.google/technology/ai/ai-creativity/

Marrone, R., Taddeo, V., & Hill, G. (2022). Creativity and Artificial Intelligence—A Student Perspective. Journal of Intelligence; 10(3):65. https://doi.org/10.3390/jintelligence10030065.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J., 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, April 2023.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, 21(4), pp. 18–21.

Nature Electronics (2023). Editorial: The year of brain–computer interfaces. *Nat Electron* 6, 643. https://doi.org/10.1038/s41928-023-01041-8.

O'Neil, C. (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Open AI (2024), Sora: https://openai.com/sora

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), pp. 447–453.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M. & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. Patterns, Volume 5, Issue 5: https://doi.org/10.1016/j.patter.2024.100988.

Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press.

Pereira, F. C. (2007). Creativity and Artificial Intelligence: A Conceptual Blending Approach. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110198560

Rafner, J., Beaty, R.E., Kaufman, J.C. et al. (2023). Creativity in the age of generative AI. Nat Hum Behav 7, 1836–1838 https://doi.org/10.1038/s41562-023-01751-1.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. Ethics and Information Technology, 20(1), pp. 5–14.

Reddy, A. (2022). Artificial everyday creativity: creative leaps with AI through critical making, Digital Creativity, 33:4, 295-313, DOI: 10.1080/14626268.2022.2138452.

Ricci RC, Paulo ASC, Freitas AKPB, Ribeiro IC, Pires LSA, Facina MEL, Cabral MB, Parduci NV, Spegiorin RC, Bogado SSG, Chociay Junior S, Carachesti TN, Larroque MM. (2022). Impacts of technology on children's health: a systematic review. Rev Paul Pediatr. Jul 6;41:e2020504. doi: 10.1590/1984-0462/2023/41/2020504. PMID: 35830157; PMCID: PMC9273128.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. New York: Viking.

Russell, S. & Norvig, P. (2020). Artificial Intelligence: A Modern Approach. 4th edn. Boston: Pearson.

Ryan-Mosley, T., Heikkilä, M. & Tng, Z. (2024), What's next for AI regulation in 2024? MIT Technology Review: https://www.technologyreview.com/2024/01/05/1086203/whats-next-ai-regulation-2024/

Schwab, K. (2017). The fourth industrial revolution. Portfolio Penguin.

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), pp. 417–457.

Shanahan, M. (2015). The Technological Singularity. Massachusetts Institute of Technology Press.

Sharkey, A. (2016). Should we welcome robot teachers? Ethics and Information Technology, 18(4), pp. 283–297.

Shtefan, A. (2021). Creativity and artificial intelligence: a view from the perspective of copyright, Journal of Intellectual Property Law & Practice, Volume 16, Issue 7, July 2021, Pages 720–728, https://doi.org/10.1093/jiplp/jpab093.

Stanford University (2023). Artificial Intelligence Index Report. Human Centred Artificial Intelligence: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy.

Tigre Moura, F. (2023) Artificial Intelligence, Creativity, and Intentionality: The Need for a Paradigm Shift. J Creat Behav, 57: 336-338. https://doi.org/10.1002/jocb.585.

University of Oxford (2024). World leaders still need to wake up to AI risks, say leading experts ahead of AI Safety Summit. New and Events: https://www.ox.ac.uk/news/2024-05-21-world-leaders-still-need-wake-ai-risks-say-leading-experts-ahead-ai-safety-summit.

Vinchon, F., Lubart, T., Bartolotta, S., Gironnay, V., Botella, M., Bourgeois-Bougrine, S., Burkhardt, J.-M., Bonnardel, N., Corazza, G.E., Glăveanu, V., Hanchett Hanson, M., Ivcevic, Z., Karwowski, M., Kaufman, J.C., Okada, T., Reiter-Palmon, R. and Gaggioli, A. (2023) Artificial Intelligence & Creativity: A Manifesto for Collaboration. J Creat Behav, 57: 472-484. https://doi.org/10.1002/jocb.597

Wallach, W. & Allen, C. (2008). *Moral Machines: Teaching Robots Right From Wrong*. Oxford: Oxford University Press.

West, S. M., Kraut, R. E. & Chew, H. E. (2019). I'd blush if I could: Closing gender divides in digital skills through education. UNESCO Report. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000367416.

Whittlestone, J., Nyrup, R., Alexandrova, A. & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 195–200.

Williamson, B. (2017). Big data in education: The digital future of learning, policy, and practice. Learning, Media and Technology, 42(2), pp. 245–247.

Wingström, R., Hautala, J. & Lundman, R. (2022). Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists, Creativity Research Journal, DOI: 10.1080/10400419.2022.2107850

Zuboff, S. (2019) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. London: Profile Books.