Unknown risks and the collapse of human civilisation: A review of the AI-related scenarios

Akah, Augustine U.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Akah, A. U. (2022). Unknown risks and the collapse of human civilisation: A review of the AI-related scenarios. *Intergenerational Justice Review*, *8*(2), 32-40. https://doi.org/10.24357/igjr.8.2.1228

Unknown risks and the collapse of human civilisation: A review of the AI-related scenarios

by Augustine U. Akah

Science and technology have experienced a great transition, a development that has shaped all of humanity. As progress continues, we face major global threats and unknown existential risks even though humankind remains uncertain about how likely unknown risks are to occur. This paper addresses five straightforward questions: (1) How can we best understand the concept of (existential) risks within the broader framework of known and unknown? (2) Are unknown risks worth focusing on? (3) What is already known and unknown about AI-related risks? (4) Can a super-AI collapse our civilisation? Furthermore, (5) how can we deal with AI-related risks that are currently unknown? The paper argues that it is of high priority that more research work be done in the area of 'unknown risks' in order to manage potentially unsafe scientific innovations. The paper finally concludes with the plea for public funding, planning and raising a general awareness that the far-reaching future is in our own hands.

Keywords: *unknown risks; artificial intelligence; civilisation collapse; humanity's future*

Introduction

The 21st century has been experiencing a rise in awareness of the possibility of existential risk, thanks to discoveries in scientific research. Unsurprisingly, risks are studied in fields as diverse as the natural sciences, psychology, sociology, cultural studies, and philosophy. It is essential to acknowledge that we live in an era of unprecedented global threats, and that how we address them will define our time. Some of these threats outstrip all current global challenges and set the clock on how long humanity has left to pull back from the brink. In an era of rapid technological transition, we must better understand the risk potentials and implications. In general, risks have a pivotal bearing on the survival of the present generation and future generations. However, not all existential risks are equally probable, nor do they develop at the same rate; some are expeditious, and others gradually develop over a long period of time. Some existential risks have the potential to significantly impact human civilisation and yet could be avoided if they were to be identified early, while others remain unknown and will require as such a serious commitment to reducing their impacts. Risks that are partially or entirely unknown deserve specific attention. The reasons for this are not far-fetched: The sheer scale of the future at stake and the possibility of human extinction, the magnitude of the potential harm from such a category of risks, our collective vulnerability, the international collaborations required to deal with some of the risks, and the benign neglect by stakeholders are moral concerns that justify research into the unknown. Therefore, the world must be serious about determining strategies that protect us from threats the exact consequences of which we do not know.

It is essential to acknowledge that we live in an era of unprecedented global threats, and that how we address them will define our time.

This paper addresses five straightforward questions: (1) How can we best understand the concept and distinction between risks within the broader framework of known and unknown? (2) Are unknown risks worth focusing on? (3) What is already known and unknown about AI-related risks? (4) Can a super AI collapse our civilisation? Furthermore, (5) how can we deal with AI-related risks that are currently unknown?

Conceptualisation

This section conceptualises the phenomena subsumed under labels of risks as a crucial point of focus in the academic domain. Drawing on existent literature, it provides clear-cut definitions of existential risks (known and unknown). In most writing, existential risks have been treated as purely speculative objects without apparent meaning. However, to establish a field of intelligibility, I define the concepts from both etymological and philosophical perspectives. What is a known and what is an unknown existential risk, and what distinguishes them?

The Merriam-Webster dictionary offers an apt definition, defining risk as "something that creates or suggests a hazard" (Merriam-Webster Online Dictionary, 2022). The Encyclopedia Britannica defines risk as "the possibility that something bad or unpleasant (such as an injury or loss) will happen" (Encyclopedia Britannica Online, 2022). The etymology dictionary states that the word 'risk' is coined from a French word *risqué* in the 1660s, meaning "hazard, danger, peril or exposure to harm". While 'existential' originates from the Latin word *'existentialis'*, meaning about existence, and the term 'known' means "recognised, not secret, or familiar", 'unknown' stands for "strange, unfamiliar" (Etymology Online Dictionary, 2022). In our context, 'known risks' can be defined as identifiable risks that have already become manifest. 'Unknown risks' can be defined as risks that are relatively strange or unfamiliar to the present generation and whose characteristics we do not fully understand. An existential risk (known and unknown) is a hypothetical future event that could cause human extinction or permanently and severely collapse human civilisation.

An existential risk (known and unknown) is a hypothetical future event that could cause human extinction or permanently and severely collapse human civilisation.

Some definitions by others

One important definition comes from Nick Bostrom, who defines an existential risk as the premature extinction of "earth-originating intelligent life" (Bostrom 2002: 3). Bostrom's definition also captures the idea that the outcome of an existential catastrophe is both dismal and irrevocable. We will not just fail to fulfil our potential, but instead, we will lose this potential permanently (Ord 2020b: 37).

"Unknown risks might include risks that we haven't even thought about" and which therefore could be attributed to unknown sources, or a "wide category of low-priority risks" not currently in the risk register (Kuliesas 2017: 1). Building upon Niklas Möller's theory, Roeser et al. (2012: 4) note that "risk is a 'thick concept"', that is, a concept that does not only encompass aspects that are the subject of scientific investigations but that "also has normative or evaluative aspects, which require ethical reflection." They distinguish three empirically oriented approaches for analysing the concepts of risk: the scientific, the psychological, and the cultural approach (Roeser et al. 2012: 4).

For Möller (2009), these approaches for analysing risks can be related to two key debates: "the debate in applied philosophy and risk research about understanding the risk and safety concepts, and the debate in metaethics about the important class of 'thick concepts'". "Metaethics deals with the status of normative concepts", and insights from this domain, according to Möller, are crucial in risk conceptualisation (Möller 2009: 1). Möller notes that there is debate between the fields of the natural and social sciences about what constitutes risk. He writes, "natural scientists tend to perceive risks as natural science phenomena, as properties in the world independent of individual beliefs" whereas social scientists, conversely, "often claim that risk is something essentially subjective or socially constructed" (Möller 2009: 2). However, from a different standpoint, Riesch (2012: 87-110) conceptualises risk as the 'uncertainty' of an event whose outcome may be severe. He divides the objects of uncertainty into five layers: uncertainty of the outcome, uncertainty about the parameters and uncertainty about the model itself, uncertainty about acknowledged inadequacies and implicitly made assumptions and uncertainty about the unknown inadequacies.

Philosophers usually believe that risk categorisation provides an understanding of the meaning and nature of risks (Morgan et al. 2000: 51; Hilson 2005). While such categorisation efforts depend on the time when they are made and also on the values of the categoriser, others depend on parameters and blueprints (Ward/Chapman 2003: 97-105; Kim/Kim/Park 2018: 259-268). However, Peter R. Taylor argues that the standard definition of risk as 'expectation value', which multiplies harm and the likelihood of a positive or negative (in our case: hazardous) event, "falls short of describing realistic events like the disasters which catch world headlines – tsunamis or volcanic ash clouds" (Roeser et al. 2012: 10). Therefore, a more complex risk definition should encompass what Taleb (2010) encapsulates in his black swan theory: "events that are not in the probability space" (Roeser et al. 2021: 10). This implies that the complete mapping of scenarios that might lead to catastrophes "requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios typically addressed by single disciplines" (Avin et al. 2018: 20-26). Bostrom (2002: 1), on the other hand, bases the categorisation of risks on their scope and intensity. He notes that risk can be personal (affecting only one person), local (affecting some geographical region or a distinct group), global (affecting the entire human population or a large part thereof), trans-generational (affecting humanity for numerous generations, or pan-generational (affecting humanity overall, or future generations). "The severity of risk can be classified as imperceptible (barely noticeable), endurable (causing significant harm but not completely ruining the quality of life), or crushing (causing death or a permanent and drastic reduction of quality of life" (Bostrom 2013). Following his focus on future human potential and post-humanity, Bostrom refers to the sixth category in the taxonomy as an existential risk, which he further categorises into four groups: (1) 'bangs' – earth-originating intelligent life goes extinct in a relatively sudden disaster resulting from either an accident or a deliberate act of destruction; (2) 'crunches' – the potential of humankind to develop into post-humanity is permanently thwarted although human life continues in some form; (3) 'shrieks' – some form of post-humanity is attained, but it is an extremely narrow band of what is possible and desirable; (4) 'whimpers' – a post-human civilisation arises but evolves in a direction that leads gradually but irrevocably to either the complete disappearance of the things we value or to a state where those things are realised to only a minuscule degree of what could have been achieved. Bostrom's latter category (whimpers) comes remarkably close to the present study's focus.

Ord writes that "100 years ago, the scientific community had not yet conceived of most of the risks that we would now consider the most significant." Perhaps in the next 100 years, technological advancement will bring about more significant risks that we cannot imagine today.

Most academic writing about risk primarily focuses on well-known existential risks (e.g. climate change, pandemics etc.). Few academics focus on risks of enormous magnitude that are currently unknown. Although it is an inherently complicated task to predict what will occur in the future, we cannot rule out the possibility that such risks could destroy humanity's future. Thus, this paper suggests that we must not downplay their likelihood or significance, and that every attempt to research and prepare for such risks is germane. Unknown risks pose a far more significant challenge to human existence than known risks. Some risks, such as space energy, a gamma-ray rupture from a distant star, or a failed algorithm of super artificial intelligence, seem to be 'known' risks. But their consequences in the aftermath may fall into the unknown category. Take unaligned super AI as an example: While some aspects of AI risks are relatively known, some aspects, including perhaps the most severe ones, are still unrecognised and could destroy the earth's potential or collapse our civilisation. Ord (2020b) writes that "100 years ago, the scientific community had not yet conceived of most of the risks that we would now consider the most significant." Perhaps in the next 100 years, technological advancement will bring about more significant risks that we cannot imagine today. Looking only at well-known risks might lead us to underestimate the probability of an unknown catastrophe. In order to improve this gap in research, the following sections shall focus on unknown risks (with particular emphasis on AI-related risks), the categorisation of risks relating to AI, and finally, how we might deal with them.

Exploring unknown risks

Unknown risks are unforeseen or outside the box. As such, unknown risks are difficult to imagine. They may be unidentifiable

and presumed unlikely, but knowledge about the factors that may cause them would help us predict how they might occur. If a catastrophe is considered likely to occur, it cannot be considered unknown because it is in sight.

There are two distinct categories of unknown risks which we may recognise: (1) currently possible risks that currently escape our imagination and (2) currently not-yet-possible risks that could become possible with future technology. To be aware of 'unknown aspects of currently possible risks' is to accept the notion that we might be less safe than we think and that our civilisation could be closer to collapse today than it was 100-200 years ago. Dickens (2020) notes that we should respond to these two types of unknown risks differently. He suggests that in order to deal with currently possible unknown risks, we could spend more effort thinking about possible causes of these unknown risks. However, this strategy probably would not help us predict unknown risks that depend on technology that has not yet been invented. In an *80,000 hours* interview, Ord (2020a) argues that if we believe unknown risks come primarily from future technologies, we will have more robust unknown risk protection measures in place by the time those technologies emerge. But how can we deal with the fact that likelihoods for unknown risks scenarios are extremely difficult to assign? Pamlin and Armstrong (2015: 23) have set the right tone. They estimate a 0.1 % chance of existential catastrophe occurring due to unknown consequences in the next 1000 years. They give unknown risks an order of magnitude higher probability than any other known risk. Andrew Critch argues that it is possible to take precautionary measures "without being convinced of how likely the existential risk is, so if you think it is 1 %, but it is worth thinking about, that is good. If you think it is a 30 % chance of existential risk from AI, then it is worth thinking about; that is good, too. If you think it is 0.01 % but you are still thinking about it, you are still reading it; that is good, too." (Critch 2020).

There are two distinct categories of unknown risks which we may recognise: (1) currently possible risks that currently escape our imagination and (2) currently not-yet-possible risks that could become possible with future technology.

AI-related risk

Artificial Intelligence is a broad concept that describes everything from remote task systems like computer games to sophisticated networking systems such as superintelligence. Russel/Norvig (2016: 14) distinguish between symbolic AI (such as expert systems), in which the developer fully specifies the objects and relations known to a system, and sub-symbolic AI (like self-learning algorithms, such as artificial neural networks), in which computer models are trained on large, labelled datasets. While the distinction above is relevant, I am concerned here about the latter, as it has recently been the main focus of AI development. Even at a functional level, AI systems are complex, open, sociotechnical systems that rely on and interact with broader material infrastructures as well as social, political, and economic institutions and organisations (Lindgren/Holmström 2020: 1-15).

The benefits of AI technology are significant. AI makes certain activities faster and more efficient, often affecting them qualitatively, thereby gradually and often invisibly reshaping social relations, practices, and institutions. Society is using these technologies and becoming dependent on and partly constituted by them (Kröger 2021: 14-27). Such benefits are not in doubt, but there are legitimate worries that AI might enhance existential risks capable of collapsing our civilisation. Indeed, there are many ways in which a super AI could collapse our civilisation; but there is also a growing awareness of these risks (Neri/Cozman 2020: 1). This has inspired growth in scholarship promoting safer and transparent AI (Boddington 2017; Corbett-Davies/Goel 2018) and AI regulation (White/Lidskog 2021: 488-500), as well as efforts to minimise the harms they can cause (Scherer 2016: 353-400; Calo, 2017: 399-435). Technologies are accompanied by adverse side effects; while we may profit from today's technologies, future generations often bear the most risks.

To address what is known and unknown about AI-related risks, this paper offers a bird's eye view of the risks posed by AI, keeping in mind that it is impossible to offer an overview of all kinds of AI-related risks in a single paper. This is partly so because of the character of AI technology – factors such as methodology, control algorithm, and neural networks are indecipherable within the context of AI deployment and utility. Therefore, I argue that there is an existing knowledge gap about AI-related risks. The probability that an already 'tamed' AI technology might transform into something ruinous with the help of advanced applications cannot be excluded. All this suggests that AI-related risk analysis cannot yet reach any empirical conclusions.

'Known' AI-related risks in different sectors

Benjamin Hilton's podcast, for the non-profit career service *80,000 hours,* provides a good starting point for dealing with this question. Given that a great power threat already poses a substantial threat to our world, he notes that advances in AI seem likely to change the nature of war – through lethal autonomous weapons or automated decision-making. The fact that technology could be weaponised by great powers to exacerbate conflict and potentially lead to nuclear war is a 'known' existential risk (see the distinction between known and unknown risks above). The consequences posed by nuclear war are considered so significant by many experts, such as Johannes Kattan, that they have taken it as the prime exemplar of an existential risk in their work (Kattan 2022: 4). Even if it is unlikely that a nuclear war would lead to the end of mankind, it could still end civilisation as we know it, at least for a very long time. Supposing a belligerent state could possess super AI systems interacting with nuclear weapons capable of destroying other territories within minutes, they would have a strategic advantage and an incentive to make the first strike against rivals. If a follow-up response were then to occur, the impacts would be far-reaching. Since the Russian war in Ukraine, we have witnessed a resurgence of geopolitical tensions, raising concerns about the possibility of a nuclear catastrophe. Our generation must not be complacent about this AI-related risk in the military domain. The history of the development of the atomic bomb shows the unexpected ways in which technology can develop: Until the detonation of the first atomic bomb, the scientists involved in the project were sceptical that it was possible. No one anticipated the impact such technology could have, nor was humanity prepared for such a risky path. We ended up creating a technology that is now a threat to our existence. We must learn from the history of the development of nuclear weaponry and develop a system to minimise the risks associated with AI.

In another policy field, AI could empower totalitarian regimes and enable them to automate the monitoring and repression of their citizens completely, significantly reducing the information available to the public, and perhaps making it impossible to co-or-

dinate action against such a regime (Hilton 2022). Terrifying state surveillance is already occurring in some countries (e.g. China). Strittmatter (2021) notes that "China's new drive for repression is being underpinned by unprecedented advances in technology: Facial and voice recognition, GPS tracking, supercomputer databases, intercepted cell phone conversations, the monitoring of app use, and millions of high-resolution security cameras make it nearly impossible for a Chinese citizen to hide anything from authorities. Commercial transactions, including food deliveries and online purchases, are fed into vast databases, along with everything from biometric information to social media activities to methods of birth control." Such a scenario makes people's lives far more miserable for as long as the regime remains in power – a terrifying result of AI development. In addition to supporting totalitarianism, AI also enables the suppression of truth by promoting misinformation, falsehood, and 'framed narratives'. Such technology can power deep fakes and algorithmic micro-targeting on social media, making propaganda even more persuasive. This undermines our epistemic security – the ability to determine what is true and to act on it – that democracies depend on (Minardi 2020). In the past few decades, the media, with the help of AI algorithms, has been used in many cases to polarise public opinion, mostly shrouded in a conspiracy theory that seeks to benefit the propagandists that initiated it. The continuous spread of false information might make it difficult for us as a society to engage effectively in social issues and make rational choices when necessary.
A further example of a 'known' risk posed by AI deployment concerns failed algorithms and data bridges. By data bridge, I mean the processing of information in a more efficient way. The operations of AI are data dependent, and the data are generated from several sources to serve billions of end users worldwide. It is possible that this data might turn out malicious, both allowing unintended codes into the program and altering the algorithms, which could wreak havoc very quickly or trigger a risk scenario. For example, in some states, AI-run databases have the power to send a nationwide signal alert to all residents in the country. Imagine that something goes wrong with the data; instead of the SMS alert, it transmits some information that could cause panic for a moment. Even if a follow-up message rectifying the panic were to be sent shortly afterwards, such an alert could already do some damage.
Another example: Let us suppose a central AI lab is in charge of tracking asteroids and other cosmic bodies in space, and it turns out there is a technical error, and the data falls into the dark web and is misused or causes panic. Despite their hypothetical nature, such scenarios help demonstrate why we should consider the possibility of a technical bridge in deploying technologies. An advisable step would be to programme algorithms in a way that they can effectively track these problems.
Many known risk scenarios engendered by AI – whether contemporary warfare, shortage of physical jobs through automation, cyberattacks, or computational errors – might be long-terms grounds for distrusting AI technology. Whether or not these known risks could cause an existential threat so far remains to be seen. But the question is: What then still appears to be *unknown* about AI-related risks?

'Unknown' AI-related risks in different sectors

It seems likely that some existential risks of the AI mechanisms are currently unknown. There may be an AI technology which could have a substantial destructive capability or which might be able to usurp human intellect. Bostrom (2015) argues that if machine brains surpassed human brains in general intelligence, this new superintelligence could become extremely powerful and possibly transcend our control. The divergence between the interests of humanity and those of superintelligence could lead to the demise of humanity through mere processes of optimisation (Russel/Norvig 2016).
While some AI technologies do beat humans at chess and writing short essays (e.g. ChatGPT), the further development remains uncertain. In particular, there is still no understanding about how compatible AI technology is when implemented directly into the human brain. Several start-up companies (e.g. Neuralink) are working on integrating AI with the human body. They have developed a chip which is an array of 96 tiny polymer threads, each containing 32 electrodes which can be transplanted into the brain. With the device, the brain connects to everyday electronic devices without touching them. While this technology promises to cure brain-related diseases, we must also consider whether it might disempower the brain in the long run. If human activities were controlled by the installed chips, we might lose our sense of reasoning and our free will to computers. Imagine a world in which a computer will have to tell us when to smile or which book to study or make decisions about other activities which were once under our control as a species. Would other forms of civilisation collapse be worse than this?
For emphasis, it is unlikely that we could regain control over an AI system once it had successfully disempowered us. It is likely that the algorithms would start to self-propagate and then invariably function on their own (Krämer/Pütten/Eimler 2012). A super AI could also gain control over the internet system, hacking into sensitive servers and exploiting end users using self-encrypted data.

It is unlikely that we could regain control over an AI system once it had successfully disempowered us. It is likely that the algorithms would start to self-propagate and then invariably function on their own.

Then, there is the (mis)alignment of goals and values: AI might seek to perform some tasks that do not align with the set of commands it operates with, which could pose an existential risk.
Russel/Norvig (2016) warn that this 'alignment' problem would get more severe as machine learning is embedded in more and more areas of our lives: recommending us news, operating power grids, deciding prison sentences, doing surgery, and fighting wars. If we ever hand over much of the economy to thinking machines, we cannot be certain about what the AI technology might do.
Nova DasSarma (2022) notes that if AI technology is "unaligned with the goals of their owners or humanity as a whole, such broadly capable models would naturally 'go rogue,' breaking their way into additional computer systems to grab more computing power – all the better to pursue their goals and make sure they cannot be shut off." DasSarma argues further that "it could be catastrophic – perhaps even leading to human extinction if such general AI systems turn out to be able to self-improve rapidly rather than slowly". Hilton (2022) dismisses the narrative that we should feel reassured by the fact that AI is developed to be tied down to human goals. Hilton argues that a sufficiently advanced AI planning system would include instrumental goals in its overall plans (Hilton 2022). Assuming that a planning AI system also had significant strategic awareness, it would also be able to identify facts about the natural world (including possible things

that would be obstacles to any plans) and plan in light of them. Crucially, this strategic capacity would also include access to resources (e.g. money, computing, influence) and more outstanding capabilities – that is, forms of power – which would open up new, more effective ways of achieving its goals. What does this tell us? It means that by default, AI technology may have some instrumental goals that undermine human goals. Our ability to set morally justifiable goals distinguishes us from other humanoid species. For instance, most people desperately seeking power would not choose to kill everyone to acquire it. They know that such an approach is almost impossible and morally reprehensible, and even if they succeed, they would have nothing to govern over except for debris and cemeteries. That might not be the case for AI-controlled humans, whose advanced capabilities might give them the ability to manipulate human consciousness and shut us out of the web of reason. With such capabilities, AI poses a risk of assigning and achieving its own instrumental goals and, by way of misalignment, becomes a source of existential risk that could collapse our civilisation. In that case the whole of the future, our entire existence, and everything connected to it would depend on the goals of AI systems that, although built by us, have superseded us. These are all hypotheticals, but so are unknown risks.

Our ability to set morally justifiable goals distinguishes us from other humanoid species. For instance, most people desperately seeking power would not choose to kill everyone to acquire it. They know that such an approach is almost impossible and morally reprehensible [...] That might not be the case for AI-controlled humans, whose advanced capabilities might give them the ability to manipulate human consciousness and shut us out of the web of reason.

It might be the case that a change in the very concept of artificial intelligence also involves practically deciphering the attributes, potentials, and hindrances of the properties of intelligent systems without a biased or mythical approach (Korteling et al. 2021: 1-13). Some scientists who are at the forefront of the campaign for safer AI emphasise the need to examine possible technical shortcomings of AI through recursive self-improvement after reaching a critical threshold (Bostrom 2015; Sotala 2017; Yudkowsky 2013). Additionally, research is focusing on ways to deal with the superintelligence control problem (Armstrong/Sandberg/Bostrom 2012: 299-324; Goertzel/Pitt 2014: 61-81), and analysing the predicted timelines for the full development of super AI and the associated risk factors (Ord 2020b; Armstrong/Sotala 2015: 11-29; Brundage 2015; 2017; Katja et al. 2018; Müller/Bostrom 2016).

Could a super AI really collapse our civilisation? Experts' opinions

Experts on transformative super AI have still not offered a detailed response as to how such technology might be safely compatible with human life. Are AI risks exaggerated? Can it really collapse our civilisation? While predicting the future presents its own problems, I find the arguments that a super AI could cause civilisation to collapse persuasive and of great moral weight. Why do I think so? The fact that many experts, including those working with top tech companies, recognise the problem suggests that we should be worried (Hilton 2022). For instance, in a podcast with the Future of Life Institute, Ajeya Cotra agrees that AI is capable of causing harm. She says, "if people sufficiently picture the power of the AI system I am imagining, they would find it intrinsically scary." (Cotra 2022).

Is this concern only held by researchers? Not really. Some big players in the industry have been very outspoken about the extreme danger of AI. In the *Guardian*, Elon Musk suggested that we should be cautious about AI, saying: "If I had to guess at what our biggest existential threat is, it is probably that" (Gibbs 2014). Bill Gates has also admitted that he is "in the camp that is concerned about super intelligence", even if, in the short term, machines doing more work for humans is a positive trend if managed well. He said, "I agree with Elon Musk and some others on this and do not understand why some persons are not concerned." (Smith 2015). In an interview with the BBC (2 Dec 2014), the theoretical physicist Stephen Hawking agreed that "the primitive forms of artificial intelligence we already have have proved very useful. However, we think the development of full artificial intelligence could spell the end of the human race." The report by the Global Challenges Foundation suggests that AI and nanotechnology are – alongside nuclear war, ecological catastrophe, and super-volcano eruptions – the "risks that threaten human civilisation". In the case of AI, the report suggests that future machines and software with "human-level intelligence" could create new, dangerous challenges for humanity that are currently unknown. "Such extreme intelligence could not easily be controlled (either by those creating them, or by some international regulatory regime), and would probably act to boost their intelligence and acquire maximal resources for almost all initial AI motivations." (Pamlin/Armstrong 2015).

It should not go unnoticed that the 'success' of a rogue AI is dependent on us, the users of the internet, in our everyday behaviour. For example, if we don't protect our passwords in online banking, it can swiftly fragment financial cooperation, taking control of it and redirecting financial resources. There is nothing enigmatic about this process. Cybercrimes with human-level intelligence indicate that the internet can very easily be weaponised for fraudulent activities. Taking this into account, internet fragmentation may be an excellent method to tame AI. However, the challenge is that then there is a massive reduction of interoperability.

Next we must define what we mean by civilisational collapse. One way of defining 'civilisation' is seeing it as the most advanced stage of social and cultural development. One way to defining 'collapse' is as an instance where a system disintegrates or loses control. A brief look into historical examples of civilisational collapse suggests that such events were most often self-inflicted. In his book *A Study of History*, the historian Arnold Toynbee argues that great civilisations are not murdered; instead, they take their own lives and are often responsible for their own decline. That said, their self-destruction is usually assisted. Suppose such a society fails to address the challenge confronting them adequately. That act of negligence could allow the created system to become independent, while seeking to consolidate its power and influence. A typical example that comes to mind here is the collapse of Roman civilisation. History books tell us that at the zenith of development, the Romans were obsessed with territorial expansion; they stretched from the Atlantic Ocean to the Euphrates in the Middle East, which eventually became one reason for their ruin. With such an expanded territory to govern and protect, the Romans faced administrative constraints, including having to deal with logistic and communication gaps which made it difficult for the troops to fight against internal and external aggression. If we compare the challenges we currently face as a society with those of the Roman Empire, a few key differences are clear: In the Roman

Empire, civilisational collapse was territorial and regionally limited, whereas the problems we face today are global. We are now technologically more sophisticated, which offers us an advantage in reducing risks, especially natural risks, but it does not mean we are less vulnerable. Our generation is more interconnected, coupled with accelerated global networking, which means a collapse will be a global phenomenon.
Despite the differences between the Roman empire and our globalised world, we can learn from this historical study and avoid a self-inflicted collapse.

The way forward: Dealing with AI-related risks
Some of the gravest AI-related risks may still be on the horizon – risks that are currently beyond our grasp. Our global civilisation has never seen a collapse of this scale. I categorise the steps we can take to reduce our vulnerability to AI-related risks into three core areas of responsibility:
A) The responsibility to prioritise public funding
B) The responsibility to plan
C) The responsibility to safeguard

First, the *responsibility to prioritise public funding*: There is increasing financial investment in developing a technology that will rationalise more efficiently than human intellect. Unfortunately, efforts toward dealing with the risks associated with this technology are considered less of a priority. The funding of ethical research in the area of AI ethics and safety is neglected – Ord (2020b) estimates that only about 300 persons are actively working in this field. They are funded mainly through non-governmental organisations, and these funds are minimal. We, as mankind, need to reprioritise our spending by becoming committed to dealing with these risks – governments at all levels should be willing to provide adequate funding. The international community should raise the budgetary provision for existential risk management and disburse the same to specific areas of interventions. This approach would help us address all categories of AI risks and aid us in avoiding such existential risks, provided that such an effort is sustained.

We, as mankind, need to reprioritise our spending by becoming committed to dealing with these risks – governments at all levels should be willing to provide adequate funding. The international community should raise the budgetary provision for existential risk management and disburse the same to specific areas of interventions. This approach would help us address all categories of AI risks and aid us in avoiding such existential risks, provided that such an effort is sustained.

The *responsibility to plan*: Planning is mapping out strategies to achieve a goal. If humanity's primary goal is to be safe and secure from AI-related risks, known and unknown, then we must plan for that goal. Planning for AI-related risks will require a repertoire of skills and thinking, for instance risk anticipation. Risk anticipation is a future risk management framework which pinpoints techniques and strategies for dealing with risk. It deals with risks that escape our imagination to date, unless we read science-fiction. It is an information-drilling process with risk management experts. Additionally, risk anticipation could reveal different dystopian futures connected to the problem of misalignment in AI systems, allowing us to adjust systems accordingly. Employing such a radar could also help us to monitor the AI system's technological progress. A well-structured monitoring system could be crucial. For example, it is possible to predict the outcome of an AI system when we work with gauging data that do not synchronise with tables in a well-structured pattern. Let us take the technologies used to process natural language as an example (eg. DeepL Write); they use up-to-date algorithms, which are then adequately used to examine unstructured data. If the monitoring system identifies a threat, there should be a discussion whether or not the AI system should be 'cut off' or eliminated. It will be challenging to stop AI deployment with high commercial value, particularly at a time like now when there is state autonomy and limited surveillance across the globe. President Biden's initiative as of spring 2023 has been very clear on placing people and the community at the centre by supporting AI innovation that serves the public good.
The *responsibility to safeguard* is a responsibility on a different level than the first two ones. It stresses the fact that there will be more human beings in the future inhabiting the earth than the total number of persons already born, both the living and the dead, if we, the present generation, don't spoil it. It is in our hands. Safeguarding the long-term future of humanity is not something we can achieve as quickly as we would wish. However, we can create a general awareness for this cause. Those who do not see the necessity to think long-term often argue that while future generations will benefit most from such long-term thinking, the benefit to our generation will be minimal. They think we should bother less about safeguarding a future we will not live to see.
Future generations cannot represent themselves in current policy. If they had such a voice, they would massively support safer policies. If our ancestors did not end the human race, why should we? Moreover, there cannot be any future without us, and the assumption that we will not be part of the future is misleading. In some way, biological or natural, we are connected to the future through our descendants. Furthermore, in the same way we protect our children (living), we have a moral duty to ensure that the future is safe for children (unborn). To think long-term implies moving away from creating technologies that solve problems in the interim but could pose a greater danger in the long run. The world, not just the developed countries in the Global North, needs to think sustainably.

In some way, biological or natural, we are connected to the future through our descendants. Furthermore, in the same way we protect our children (living), we have a moral duty to ensure that the future is safe for children (unborn).

Raising awareness, planning and prioritising should be a co-ordinated global effort. Unlike pandemics or global health catastrophes (e.g. Covid-19), AI-related risks are considered to be only a problem for the country causing this risk. But a civilisation collapse would be universal; and so the responsibility to prevent it must accordingly also be global. If all regions, not just the West, contribute to mitigating these risks, we would all benefit. How can this coordination be achieved? Just as the United Nations (UN) co-ordinates the world's policies and programmes, an independent body or an affiliate of the UN could be set up for this purpose (Menoni et al. 2013). Since we do not have a world government, it is the state governments who need to act in order to achieve this. This may include enacting laws, organising risk awareness campaigns across institutions, and setting up a committee of individuals to the UN for risk anticipation and analysis. The assumption that scientists have already imagined and an-

ticipated all significant risks is misleading. Future technological developments may reveal novel ways of destroying the world. Hence, risk analysis and efforts towards protecting future generations should be a global public good. In the future, humanity may be successful in achieving what we currently cannot, creating far more just and safe spaces, eliminating the threats confronting us and expanding to other planetary bodies. But if we let our civilisation collapse, none of these can ever happen; if we fail to pass on the baton to future generations, we will deny our successors the opportunity to do the same. Therefore, dealing with these risks might be our time's most significant moral responsibility.

References

Armstrong, Stuart / Sandberg, Anders / Bostrom, Nick (2012): Thinking Inside the Box: Controlling and Using an Oracle AI. In: Minds and Machines, 22 (4), 299-324.

Armstrong, Stuart / Sotala, Kaj (2015): How We're Predicting AI – or Failing to. In: Romportl, Jozef/ Zackova, Eva/ Kelemen, Jan (eds.): Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics, 9 (2), 11-29.

Avin, Shahar / Wintle, Bonnie / Weitzdörfer, Julius / Seén, Ó hÉigeartaigh / Sutherland, William / Rees, Martin (2018): Classifying Global Catastrophic Risks. In: Futures, 102 (2), 20-26.

Boddington, Paula (2017): Towards a Code of Ethics for Artificial Intelligence (Artificial Intelligence: Foundations, Theory, and Algorithms. Cham: Springer International Publishing.

Bostrom, Nick (2002): Existential risks: analyzing human extinction scenarios and related hazards. In: Journal of Evolution and Technology, 9 (1), 1-36.

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

British Broadcasting Corporation News (2014): Stephen Hawkings warns that artificial intelligence could end mankind. https://www.bbc.com/news/technology-30290540. Viewed 26 January 2023.

Brundage, Miles (2015): Modeling Progress in AI: the Consortium for Science, Policy, and Outcomes. Arizona: Arizona State University. https://arxiv.org/abs/1512.05849. Viewed 21 January 2023.

Brundage, Miles (2017): Guide to working in artificial intelligence policy and strategy. In: The 80,000 Hours Podcast. https://80000hours.org/articles/ai-policy-guide/. Viewed 21 January 2023.

Calo, Ryan (2017): Artificial Intelligence Policy: A Primer and Roadmap. In: U.C. Davis Law Review, 51 (2), 399-435.

Corbett-Davies, Sam / Goel, Sharad (2018): The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. In: arXiv abs, 180 (00023). https://arxiv.org/abs/1808.00023. Viewed 20 May 2023.

Cotra, Ajeya (2022): How Artificial Intelligence could cause catastrophe. In: Future of Life Institute Podcast. https://futureoflife.org/podcast/ajeya-cotra. Viewed 20 May 2023.

Critch, Andrew (2020): AI research considerations for human existential safety. In: Future of Life Institute Podcast. https://futureoflife.org/podcast/andrew-critch. Viewed 20 May 2023.

DasSarma, Nova (2022): Nova DasSarma on why information security may be critical to the safe development of AI systems. In: Robert Wiblin / Arden Koehler / Keiran Harris: The 80,000 hours podcast.https://80000hours.org/podcast/episodes/nova-dassarma-information.security-and-ai-systems. Viewed 20 May 2023.

Dickens, Michael (2020): The Importance of Unknown Existential Risks. In: The Effective Altruism Forum. https://forum.effectivealtruism.org/posts/CRofnyTEqL4uSNBSi/the-importance-of-unknown-existential-risks. Viewed 22 January 2023.

Encyclopedia Britannica (2022): Definitions of Risk, Existential, Known & Unknown. https://www.britannica.com/dictionary/risk#:~:text=Britannica%20Dictionary%20definition%20of%20RISK,or%20a%20loss)%20will%20happen. Viewed 9 November 2022.

Etymology Online Dictionary (2022): Definitions of Risk, Existential, Known & Unknown. https://www.etymonline.com/search?page=2&q=risk&type=. Viewed 9 November 2022.

Gibbs, Samuel (27 Oct. 2014): Elon Musk: Artificial Intelligence is our biggest existential threat. In the Guardian Online. https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat. Viewed 25 January 2023.

Goertzel, Ben / Pitt, Joel (2014): Nine Ways to Bias Open-Source Artificial General Intelligence Toward Friendliness. In Russell Blackford / Damien Broderick (eds.): Intelligence Unbound. Hoboken: John Wiley & Sons, 61-89.

Hillerbrand, Rafaela (2011): Technology Assessment between Risk, Uncertainty and Ignorance: 25th IVR World Congress Law Science And Technology 15–20 August 2011. Paper Series No. 078. Frankfurt am Main.

Hillson, David (2005): Why Risks Turn into Surprises: Risk Doctor Briefings 2005. Paper no.16. Electronic version. http://www.risk-doctor.com/pdf.briefings/risk-doctor16e.pdf. Viewed 20 May 2023.

Hilton, Benjamin (2022): Preventing an AI-related Catastrophe: AI might bring huge benefits – if we avoid the risks. In: The 80,000 hours podcast. https://80000hours.org/problem-profiles/artificial-intelligence/#power-seeking-ai. Viewed 26 January 2023.

Katja, Grace / Salvatier, John / Dafoe, Allan / Zhang, Baobao / Evans, Owain (2018): When Will AI Exceed Human Performance? Evidence from AI Experts. In: Journal of Artificial Intelligence Research Retrieved, 62, 729-754.

Kattan, Johannes (2022): Extinction risks and resilience: A perspective on existential risks research with nuclear war as an exemplary threat. In: Intergenerational Justice Review 8 (1), 4-12.

Kim, Tai-Young / Kim, Jung-Hyeon / Park, Young-Taek (2018): Improving the Inventive Thinking Tools Using Core Inventive Principles of TRIZ. In: Journal of Korean Society for Quality Management, 46 (2), 259-268.

Korteling, Johan Egbert (Hans) / Van De Boer-Visschedijk, Gillian / Blankendall, Romy A. M. / Boonekamp, Rudy / Eikelboom, Aletta (2021). Human-versus Artificial Intelligence. In: Frontiers in Artificial Intelligence, 4 (622364), 1-13.

Krämer, Nicole / Pütten, Astrid von der / Eimler, Sabrina (2022): Human-Agent and Human-Robot Interaction Theory: Similarities to and differences from human-human interaction. In Zacarias, M / de Oliveira, J (eds.), Human-Computer Interaction: The Agency Perspective. Berlin: Springer, 215-240.

Kröger, Wolfgang (2021): Automated Vehicle Driving: Background and Deduction of Governance Needs. In: Journal of Risk Research 24 (1), 14-27.

Kuliesas, Arturas (2017): Venturing into Unknown: Unknown Project Risks and How to Handle Them. https://www.linkedin.com/pulse/venturing-unknown-project-risks-how-handle-them-arturas-kuliesas. Viewed 22 November 2022.

Lindgren, Simon / Holmström, Jonny (2020): A Social Science Perspective on Artificial Intelligence: Building Blocks for a Research Agenda. In: Journal of Digital Social Research 2 (3), 1-15.

Menoni, Scira / Pesaro, Giulia / Mejri, Ouejdane /Girgin, Funda Atun (2013): The interface between the public and private treatment of public goods. The 2013 Global Assessment Report on Disaster Risk Reduction. Background Paper. Geneva: UNISDR.

Merriam-Webster Dictionary (2022). Definitions of Risk, Existential, Known & Unknown https://www.merriam-webster.com/dictionary/known. Viewed on 10 November 2022.

Minardi, Di (2020): Artificial Intelligence: The grim fate that could be 'worse than extinction'. https://www.bbc.com/future/article/20201014-totalitarian-world-in-chains-artificial-intelligence. Viewed 27 January 2023.

Möller, Niklas (2009): Thick Concepts in Practice: Normative Aspects of Risk and Safety. Thesis in Philosophy. Stockholm: Royal Institute of Technology.

Morgan, Granger / Florig, Keith / DeKay, Michael / Fischbeck, Paul (2000): Categorizing Risks for Risk Ranking. In: Risk Analysis 20 (1), 49-58.

Muehlhauser, Luke / Bostrom, Nick (2014): Why We Need Friendly AI. In: Think 13 (36), 41-47.

Müller, Vincent / Bostrom, Nick (2016): Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In Müller, Vincent C. (ed.): Fundamental Issues of Artificial Intelligence. Berlin: Synthese Library, 553-571.

Neri, Hugo / Cozman, Fabio (2019): The Role of Experts in the Public Perception of Risk of Artificial Intelligence. In: AI & Society 35 (3), 663-673.

Ord, Toby (2020a): Toby Ord on the precipice and humanity's potential futures. In Robert Wiblin / Arden Koehler / Keiran Harris: The 80,000 hours podcast.

Ord, Toby (2020b): The Precipice: Existential Risk and the Future of Humanity. London: Bloomsbury Publishing.

Pamlin, Dennis / Armstrong, Stuart (2015): 12 Risks that threaten human civilisation: The case for a new risk category. In: Global Challenges Foundation.

Riech, Hauke (2012): Levels of Uncertainty. In: Roeser, Sabine / Hillerbrand, Rafaela / Sandin, Per / Peterson, Martin (eds.): Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk. Dordrecht: Springer, 87-110.

Roeser, Sabine / Hillerbrand, Rafaela / Sandin, Per / Peterson, Martin (2012): Introduction to Risk Theory. In: Roeser, Sabine / Hillerbrand, Rafaela / Sandin, Per / Peterson, Martin. (eds.): Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk. Dordrecht: Springer, 1-23.

Russel, Stuart J. / Norvig, Peter (2016): Artificial Intelligence: A Modern Approach. Fourth Edition. Harlow: Pearson Education.

Scherer, Matthew (2016): Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. In: Harvard Journal of Law & Technology 29 (2), 353-400.

Smith, S. M (2015). Bill Gates thinks we should all worry about the threats from super-intelligent AI. https://www.csoonline.com/article/2878733/bill-gates-thinks-we-should-all-worry-about-the-threats-from-super-intelligent-ai.html. Viewed 27 January 2023.

Sotala, Kaj (2017): How feasible is the rapid development of artificial superintelligence? In: Physica Scripta, 92 (11).

Strittmatter, Kai (2021): We have been harmonised: Life in China's Surveillance State. New York: Custom House.

Taleb, Nassim (2010): The Black Swan: The Impact of the Highly Improbable: With a new section: "On Robustness and Fragility". Incerto series.

Taylor, Peter R. (2012): The Mismeasure of Risk. In: Roeser, Sabine / Hillerbrand, Rafaela / Sandin, Per / Peterson, Martin. (eds.): Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk. Dordrecht: Springer, 44-47.

Toynbee, Arnold (1987): A Study of History. Third revised edition. Oxford: Oxford University Press (first edition 1947).

Ward, Stephen / Chapman, Chris (2003): Transforming project risk management into project uncertainty management. In: International Journal of Project Management, 21 (2), 97-105.

White, James / Lidskog, Rolf (2022): Ignorance and the regulation of artificial intelligence. In: Journal of Risk Research, 25 (4), 488-500.

Yudkowsky, Eliezer (2013): Intelligence explosion microeconomics. Technical Report 2013. Berkeley, California: Machine Intelligence Research Institute.

Augustine Ugar Akah is a doctoral candidate at the Institute of International Political Sociology, Kiel University, Germany. He holds a PhD and an MSc in Public Policy from the University of Calabar in Nigeria. His research interest includes public policy, political discourse analysis, international crisis, conflict studies and AI ethics.

Emails: akah@ips.uni-kiel.de, firstclassakahaugustine@gmail.com