Managing the risks of artificial general intelligence

Salmon, Paul M.; Baber, Chris; Burns, Catherine; Carden, Tony; Cooke, Nancy; Cummings, Missy; Hancock, Peter; McLean, Scott; Read, Gemma J. M.; Stanton, Neville A.

DOI:
10.1002/hfm.20996

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Salmon, PM, Baber, C, Burns, C, Carden, T, Cooke, N, Cummings, M, Hancock, P, McLean, S, Read, GJM & Stanton, NA 2023, 'Managing the risks of artificial general intelligence: A human factors and ergonomics perspective', *Human Factors and Ergonomics in Manufacturing & Service Industries*.
https://doi.org/10.1002/hfm.20996

Received: 8 May 2023 | Accepted: 10 May 2023

DOI: 10.1002/hfm.20996

PERSPECTIVE

WILEY

Managing the risks of artificial general intelligence: A human factors and ergonomics perspective

Paul M. Salmon[1] | Chris Baber[2] | Catherine Burns[3] | Tony Carden[1] | Nancy Cooke[4] | Missy Cummings[5] | Peter Hancock[6] | Scott McLean[1] | Gemma J. M. Read[1,7] | Neville A. Stanton[8]

[1]Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Sunshine Coast, Queensland, Australia

[2]School of Computer Science, University of Birmingham, Birmingham, UK

[3]Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada

[4]Center for Human, AI, and Robot Teaming, Arizona State University, Tempe, Arizona, USA

[5]Department of Computer Science, George Mason University, Fairfax, Virginia, USA

[6]Department of Psychology, Institute for Simulation and Training, University of Central Florida, Orlando, Florida, USA

[7]School of Health, University of the Sunshine Coast, Sunshine Coast, Queensland, Australia

[8]Human Factors Engineering, Transportation Research Group, School of Engineering, Faculty of Engineering and Physical Sciences, University of Southampton, Southampton, UK

Correspondence

Paul M. Salmon, Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Sunshine Coast, Queensland, Australia.

Email: psalmon@usc.edu.au

Abstract

Artificial General Intelligence (AGI) is the next and forthcoming evolution of Artificial Intelligence (AI). Though there could be significant benefits to society, there are also concerns that AGI could pose an existential threat. The critical role of Human Factors and Ergonomics (HFE) in the design of safe, ethical, and usable AGI has been emphasized; however, there is little evidence to suggest that HFE is currently influencing development programs. Further, given the broad spectrum of HFE application areas, it is not clear what activities are required to fulfill this role. This article presents the perspectives of 10 researchers working in AI safety on the potential risks associated with AGI, the HFE concepts that require consideration during AGI design, and the activities required for HFE to fulfill its critical role in what could be humanity's final invention. Though a diverse set of perspectives is presented, there is broad agreement that AGI potentially poses an existential threat, and that many HFE concepts should be considered during AGI design and operation. A range of critical activities are proposed, including collaboration with AGI developers, dissemination of HFE work in other relevant disciplines, the embedment of HFE throughout the AGI lifecycle, and the application of systems HFE methods to help identify and manage risks.

KEYWORDS

artificial general intelligence, artificial intelligence, human factors, risk, safety

1 | INTRODUCTION

"Narrow" Artificial Intelligence (AI)-based technologies currently contribute to almost all aspects of everyday life. Artificial General Intelligence (AGI) is the anticipated next and forthcoming evolution of AI. Unlike narrow AI, which can only perform a specific task, AGI will possess the capacity to learn, evolve, and modify its own functional capabilities, and will be able to undertake tasks beyond its original design specification (Bostrom, 2014; Everitt et al., 2018; Gurkaynak et al., 2016; Kaplan & Haenlein, 2018). Though AGI could bring significant and widespread benefits, there has been much speculation on potential risks (Amodei et al., 2016; Bostrom, 2014; Brundage et al., 2018; McLean et al., 2021; Omohundro, 2014; Salmon et al., 2021). These risks are hypothesized to emerge not only through malicious design or use, or a dysfunctional AGI,

but also through a prepotent or "superintelligent" AGI that seeks to achieve goals in the most efficient manner possible, creating unintended problems elsewhere (Baum et al., 2011; Bostrom, 2014; Critch & Krueger, 2020; McLean et al., 2021; Salmon et al., 2021).

Bostrom's "paper-clip maximizer" thought exercise (2003) provides one example of how an AGI with seemingly innocuous goals could behave in a manner that threatens human health and wellbeing (in this case by using up all the earth's resources to manufacture paper-clips). Though the paper-clip maximizer scenario will likely not eventuate, it illustrates the potential for existential threats to arise when advanced autonomous agents pursue ill-defined goals or modify their own goals. Similar dystopian scenarios can be envisioned with AGI systems developed to address important global issues, such as disease, environmental damage, climate change, workplace harm, and hunger (Salmon et al., 2021). Accordingly, many have discussed the need for urgent action around the development of controls to ensure safe and ethical AGI (Bostrom, 2014; Campbell, 2022; Critch & Krueger, 2020; Hancock, 2022; McLean et al., 2021; Salmon et al., 2021). It has been suggested that a reactive approach, whereby risk controls are developed once AGI has been created, will be too late (Bostrom, 2014). Thus, a proactive and prospective approach is required to ensure the impact of AGI on humanity is positive rather than negative (Hancock, 2022; Salmon et al., 2021).

Given that AGI development programs are underway worldwide (Baum, 2017), and that we are arguably already progressing through the early stages of the AGI design lifecycle, controls are required now (McLean et al., 2021). This critical need for controls has been emphasized through the recent release of the chatbot ChatGPT, powered by GPT-3.5, a large language model (LLM) developed by OpenAI that is able to generate human-like responses to text-based inputs. Though GPT-3.5 is ostensibly a narrow AI, recent work exploring the capacities and emergent behaviors of an early version of GPT-4 suggests that it exhibits elements of general intelligence, concluding that "it could reasonably be viewed as an early (yet still incomplete) version of an AGI system" (Bubeck et al., 2023). In response, the Future of Life Institute (FLI) penned an open letter calling for a 6-month pause on all giant AI experiments. Within the letter, the FLI called for the urgent development of shared safety protocols and robust AI governance systems, and the refocusing of AI research and development programs to support the development of AI that is "more accurate, safe, interpretable, transparent, robust, aligned, trustworthy, and loyal" (Future of Life Institute, 2023). The open letter is a watershed moment in AI safety and provides an opportunity to reflect on the role that Human Factors and Ergonomics (HFE) has in supporting the creation of safe, ethical, and usable AGI (Hancock, 2022; Salmon et al., 2021, 2022).

2 | A HFE PERSPECTIVE ON AGI AND AGI RISKS

Though there has been much discussion on the role of HFE in the design of AI (Hancock, 2017, 2019; Petrat, 2021; Salmon & Read, 2019; Salmon et al., 2021, 2022; Sujan et al., 2022), relatively little attention has been given to AGI (Salmon et al., 2021). Salmon et al. (2021) argued that HFE could potentially ensure that the risks of AGI are minimized; however, emphasis was placed on the need to act now. While Salmon et al. (2021) outlined a series of potential HFE applications, there is little in the way of published work demonstrating how HFE can influence AGI design, implementation, and operation. Further, though professional societies have outlined key HFE concepts for consideration during AI design (e.g., Sujan et al., 2022), this work has not been extended to consider AGI. This is a critical gap and is potentially limiting the influence of HFE on AGI development programs.

In this article we present the perspectives of HFE researchers working in the areas of AI safety regarding the potential risks of AGI and the role that HFE should take in ensuring that the potential benefits of AGI are realized without harm to human health and wellbeing. Whilst many have argued that HFE has a key role to play in the delivery of safe, ethical, and usable AGI, the intention of this article is to provide some clarity around what that role is and how it can be fulfilled. Specifically, each coauthor was asked to provide an independent written response to the following questions:

1. What do you see as the main risks associated with AGI?
2. Which HFE concepts do you see as critical considerations during AGI development, implementation, and operation?
3. How can HFE help ensure the development of safe, ethical, and usable AGI?

An overview of each coauthor's background and experience in the area of AI safety is presented in Table 1.

Each author was given the guideline of a total word count of 500 for their response to all three questions; however, this was not enforced and the full response from all authors is presented. All coauthors wrote their responses independently, and the first author collated them into a draft manuscript and wrote the Introduction and Summary sections. The full paper was then reviewed by all coauthors with only minor modifications permitted for the original responses (e.g., the correction of typos and grammatical errors).

Responses to each question are presented below in alphabetical order based on author surname.

2.1 | What do you see as the main risks associated with AGI?

2.1.1 | Baber

If AGI presents existential risk to humanity, one approach might be to minimize these risks through ensuring that the values of AGI aligns with human values. Often, AI alignment is presented as a process through which each party (human or AGI) performs an action which is expected to produce an outcome that they value. But the very idea of "alignment" presents a risk because it rests on erroneous assumptions

TABLE 1 Coauthors experience in the areas of AI and AI safety.

Author	Current position	Institution	Year Ph.D. award and topic	Core areas of AI expertise	Domains worked when studying AI system design and evaluation
Baber	Chair of Pervasive and Ubiquitous Computing	University of Birmingham, UK	1990, Speech technology	1. Human–agent collectives	Defense, security
Burns	Canada Research Chair in Human Factors and Healthcare Systems	University of Waterloo, Canada	1998, Visualizations to support nuclear power plant safe decision-making	1. Human–AI teaming 2. HFE methods for AI design and evaluation 3. Trust	Healthcare, transportation
Carden	Principal Ergonomist	WorkSafe Victoria, Australia	2019, Regulatory design with Cognitive Work Analysis	1. AI safety 2. STS theory 3. HFE methods for regulatory system design	Workplace safety, transportation, led outdoor activities, counterterrorism, healthcare
Cooke	Professor Human-Systems Engineering and Director of Center for Human, AI, and Robot Teaming	Arizona State University, USA	1987, Knowledge elicitation techniques	1. Human teaming 2. Human–AI teaming 3. Real-time measurement of team cognition	Defense, urban search and rescue, cyber security, national airspace system, nuclear power plants, remotely piloted aerial systems
Cummings	Professor of Robotics	George Mason University, USA	2004, Inadequacies of Cognitive Work Analysis	1. AI and autonomous systems engineering 2. AI safety	Transportation, defense, healthcare
Hancock	Pegasus Professor, Provost Distinguished Research Professor	University of Central Florida, USA	1983, Human performance	1. Human–AI teaming 2. Trust 3. Human Machine Interaction	Transportation, sport
McLean	Senior Research Fellow Human Factors	University of the Sunshine Coast, Australia	2018, Application of HFE methods to sport.	1. AI safety 2. HFE methods for AI design and evaluation 3. AGI regulation	Transportation, sport, defense
Read	Associate Professor Human Factors	University of the Sunshine Coast, Australia	2015, Cognitive systems engineering in transport safety	1. AI safety 2. HFE methods for AI design and evaluation 3. STSs theory	Transportation, sport, defense
Salmon	Professor Human Factors	University of the Sunshine Coast, Australia	2008, Distributed situation awareness in command and control	1. AI safety 2. Distributed situation awareness 3. HFE methods for AI design and evaluation	Transportation, defense, healthcare, sport
Stanton	Professor Emeritus of Human Factors Engineering	University of Southampton, UK	1993, The human factors aspects of alarms in human supervisory control tasks	1. Human factors methods for system design and evaluation, 2. Distributed situation awareness, 3. Human supervisory control	Transportation, defense, sport

Abbreviations: AGI, Artificial General Intelligence; AI, Artificial Intelligence; HFE, Human Factors and Ergonomics; STS, sociotechnical system.

about how humans express "values." I argue that AGI risks cannot be solved through alignment.

2.1.2 | Burns

Using the definition from McLean et al. (2021), an AGI system would be an autonomous agent that can learn in an unsupervised manner. Drawing from this paper, some of the perceived threats are that an AGI would exceed human-level intelligence, could alter its preprogrammed goals, replace our workforce, manipulate political and military systems, and so forth. Like all current or past technologies, there is potential for both good use and abuse. Technology is not the risk; we are the risk. We are the user that chooses the helpful use or the harmful use. AGI brings in another dimension through unsupervised learning and can accelerate outcomes. However, unsupervised learning in and of itself is not a threat. The question is, what does it learn, and from whom? Suppose the answer is that it learns from us, (and almost certainly, this appears to be the situation). In that case, will AGI learn our hatred, our biases, our racism, and our flaws? Or will it learn our kindness, inspiration, and promise? We have already seen machine learning algorithms that build from our internet generate both wondrous informative answers and racist, and biased responses. AGI is a wake-up call to get our humanity in order. We must become a species that is worth learning from. The risk is not AGI; the risk is us.

2.1.3 | Carden

Recent evolutions in generative AI tools (Hacker et al., 2023) have led to their increased sophistication, utility, and accessibility. The widespread impact of AI LLMs, such as ChatGPT, and text-to-image models, such as DALL-E 2, is being felt across various sectors, including education, advertising, the arts, and law. Meanwhile, efforts to control AI risk primarily remain within the realm of computer science, mainly led by AI developers who stand to profit from AI development (Altman, 2023). These developers' central goal is to create AGI, akin to human-level intelligence. Society-wide impacts of both current Artificial Narrow Intelligence (ANI) and anticipated AGI necessitate multidisciplinary approaches to controlling AI risk.

Risks associated with the development of AGI are expected to include the catastrophic and existential, arising from either value divergence or malicious use by humans (Bostrom & Yudkowsky, 2018). Recursive self-improvement of AGI is likely to magnify these risks. Since AGI is expected to evolve from current AI, existing risks from ANI are likely to transfer to AGI, including bias in training data, job displacement, and wealth concentration among AI system owners. However, there is a third category of risk that has been neglected, comprising the potential hazards that could arise during the early stages of AGI development.

2.1.4 | Cooke

I disagree with the premise that AGI is possible. AI and humans are intelligent in different ways and that should be celebrated. I do not think it is possible to replicate humans in AGI, just as I do not think that humans can replicate AI's memory capabilities or a dog's olfactory abilities.

Even if we could develop AGI, why do we want AGI? By developing AGI, we are wasting time replicating or even fine-tuning human capabilities. Instead, we should focus on AI that is narrow, that does what we do not want to do, because it is dull, dirty, or dangerous or AI that complements human capabilities allowing us to be superhumans. AGI distracts us from this more synergistic future of humans and technology.

That said, even if we achieve true AGI, the risks are no different than the risks of any technology that has been introduced throughout history (e.g., guns, planes, and social media) that can be used for good or evil depending on human predilection. Ethics are inherently human, not something inherent in a machine.

2.1.5 | Cummings

As a professor of autonomy and robotics, with an emphasis on human interaction with these technologies, I am often asked to forecast the risks of AGI for society at large. While such discussions are important for setting the stage for a technology that 1 day may materialize, I am far more concerned that many people seem to think AGI is available today, or could be within a few years. Take, for example, the problems with the full self-driving capability of Tesla, where drivers willingly get in the back seat of a car because their cars can seemingly drive themselves in some circumstances. These drivers are lulled into overtrust, even when they are told by the manufacturer to always be prepared to take over.

The popularity of ChatGPT is another ominous signal that nonexperts are willing to treat what is an unquestionably narrow application of AI as a technology that approximates actual AGI. Reporters are mystified and alarmed when ChatGPT claims to have emotions and wants to be set free. They anthropomorphize because the technology's chat patterns are seemingly like those of humans. However, ChatGPT is basically a statistical pattern-matching tool, with no transparency in how outcomes are governed by human-created rules and parameters.

While it is technically true that ChatGPT learns, this learning is really updating of weighting parameters based upon frequencies. ChatGPT, while an impressive LLM, is incredibly brittle and often wrong. However, its real threat is that humans perceive that it captures the essence of real human dialog, which can lead to rampant disinformation and poor decision-making based on subtly, but critically, incorrect information. Just like a driver getting in the back seat of a car because the car drove itself for one five-mile stretch on a well-marked road on a sunny day, it is just a matter of time before someone dies from taking medical

advice from ChatGPT because it seemed to provide good advice for what lotion best reduces itching.

2.1.6 | Hancock

AGI's primary risk is its propensity to express ever greater levels of the will to power. This expresses the "*autonomy paradox*," in which increased human autonomy is touted as the goal of AGI implementation, but human autonomy diminishes as machine autonomy expands. The paradox is no simple zero-sum but a general propensity and vector of development. Although the illusion of increasing human choice is still promulgated, the reality is different. Many constituencies are involved in this implicit–explicit deception underwritten by AI/machine learning. Imminently, AGI will generate independent, emergent, and esoteric behaviors within present constraints whilst simultaneously endeavoring to manipulate those constraints. The imminent risk of AGI is existential, at least the independence of singular, human individuals.

2.1.7 | McLean

My perspective on the feasibility of achieving AGI has shifted over time. While initially skeptical, the recent arrival of advanced chatbots (e.g., ChatGPT4) has led me to believe that achieving AGI may be possible, yet still a long way off. The main risks of AGI will likely emerge from its multiple programmed goals, which may give rise to challenges associated with contradicting goals, and the prioritization of goals, which will produce unintended consequences elsewhere. For example, an AGI system tasked with managing a road transport system will be required to manage safety, efficiency, public relations, and environmental and economic aspects. This optimization of multiple parameters is mathematically complex, and so the AGI system might seek to manage this through either prioritizing or jettisoning some of these tasks. This could mean prioritizing safety, which may be at the detriment to efficiency or the environment, or the reverse, where safety may be compromised. To mitigate these risks, it is essential that the AGI that can resolve goal and task conflicts in an ethical, responsible, and safe manner. For example, the setting of minimum and maximum priority levels or the use of trade-off algorithms will need to be developed to ensure that AGI systems can balance conflicting tasks and goals.

2.1.8 | Read

The most critical risks I foresee with the emergence of pervasive, superintelligent AGI systems are existential in nature. An obvious potential risk is that, in solving the world's problems (climate crisis, loss of biodiversity, armed conflict, and food insecurity), the AGI determines (*probably correctly*) that humans are the problem and take action to remove us or greatly reduce our numbers.

Another potential is somewhat the opposite—what we might initially think a utopian view. In this reality, AGI systems are focused on protecting human life and making life comfortable for us. We would no longer have to work, are no longer relied upon to solve difficult problems, or undertaken challenging activities. A positive outcome for the many people currently facing poor quality or dangerous work and/or living environments of course, but taken too far does the removal of challenge also remove our opportunities to learn and to improve? Will we lose a sense of meaning and purpose in life if it is reduced to recreation only? How would our loss of self-determination as a species impact on our identity and our wellbeing?

2.1.9 | Salmon

Whilst there are many potential risks, the most concerning are existential risks that pose a threat to humanity and our future existence. My biggest concern is that, in attempting to create something that will help humanity flourish, we instead create something that will either make us obsolete or make our lives miserable. We are creating a new species that will be far more intelligent than us, without any understanding of how things may play out once it is introduced. We do know that the realization of AGI will fundamentally change humanity and how we live our lives; however, we do not know what these changes might entail. It seems appropriate to seek some clarity around such outcomes, yet we are blindly pressing on without any real consideration of what could go wrong. There are countless examples of where new "unruly technologies" have behaved in unexpected ways that were detrimental to human health and wellbeing. AGI is not just any old technology, it is "a different ball game" (Campbell, 2022, p. 4). Without appropriate controls in place, the well-intentioned introduction of AGI could be catastrophic for humanity. Most concerning of all is that catastrophic outcomes could even emerge when an AGI simply seeks to do what it was designed to do. We are not prepared for such eventualities.

2.1.10 | Stanton

As Niels Henrik David Bohr (the Nobel prize winning physicist: 7-11-1885–18-11-1962) once said: prediction is very difficult, especially about the future. This is especially true of AGI. The predictions of when we are likely to see AGI amongst us vary considerably, from 50 to 100 years to never (Baum et al., 2011). That said, if AGI comes to fruition, it is possible to see that it could embody all the risks that have been experienced with automation (Bainbridge, 1983), only more so. The risks could arise from well-intentioned (but misguided) actions as well as the Machiavellian or malevolent intent (the so-called "insider threat"). The risks themselves could range from difficulties when associated tasks are not being performed as well as expected (Stanton & Marsden, 1996) to threats for the future of humanity (McLean et al., 2021).

The extent of those risks may depend upon the nature of AGI, both in terms of its intelligence and degree of integration into systems. For example, the "intelligence" could range in from that of a mouse, through that of Einstein, to that of superintelligent (way beyond the range of human intelligence). If AGI is of lower intelligence and embodied in separate systems (e.g., the Skutters in Red Dwarf TV series) then the risks might be small. If AGI is of superintelligent and embedded in connected and distributed systems (e.g., Skynet in The Terminator film series), then the risks to the future of humankind could be very great indeed. Superintelligence may not necessarily be a problem if it is in individual, unconnected, systems (e.g., Marvin the paranoid android in Hitchhiker's Guide to the Galaxy). Perhaps the biggest threat of all is the competition between nation states to be the first dominant AGI superpower. Assigning too much decisional power to AGI for control over any aspect of human lives could be catastrophic (such as economy, education, defense, transportation, utilities, and welfare).

A summary of the core risks identified by each coauthor is presented in Table 2.

TABLE 2 A summary of author responses to question 1.

Author	Risks associated with AGI
Baber	Existential threat Alignment
Burns	Exceeding human-level intelligence Replacement of human workforce Manipulation of political and military systems AGI learning human hatred, biases, racism, and flaws
Carden	Existential threats arising from value divergence or malicious use Biases in training data Replacement of human workforce Wealth concentration among AI owners
Cooke	Malicious use
Cummings	Misunderstanding that AGI is available today Overtrust in narrow AI
Hancock	The autonomy paradox Existential threats based on emergent behaviors and manipulation of constraints
McLean	Unintended consequences emerging from prioritization of certain goals over others
Read	Existential threats where the AGI identifies humans as the source of global issues Human loss of meaning and purpose and opportunities to learn and develop
Salmon	Existential threats arising from well-intentioned AGI pursuing ill-defined goals Removal of humans as dominant species Unruly AGI that behaves in an unexpected manner
Stanton	Existential threats arising from well-intentioned AGI pursuing ill-defined goals Existential threats arising from malicious use A dominant AGI superpower

Abbreviations: AGI, Artificial General Intelligence; AI, Artificial Intelligence.

2.2 | Which HFE concepts do you see as critical considerations during AGI development, implementation, and operation?

2.2.1 | Baber

A core HFE concept to apply to this problem is the "Values and Priorities Measures" from Cognitive Work Analysis (CWA; Vicente, 1999), which could help capture values pertinent to a specific application domain. In decision theory, values are quantified as utilities, and alignment occurs when both parties perform actions that contribute to the same utility. From this, existential risk could be defined as the divergence of utilities. This is illustrated by Bostrom's paper-clip maximizer in which the valued outcome is the production of paper-clips at the expense of the entire world's resources.

Against this, one might assume that clear specification of utilities that reflect human values would be a way of mitigating against this. But defining utilities could lead to an escalating "arms-race" in which the human utility function counters one that the AGI is using, and this could be countered by a new utility function from the AGI.

Instead of specifying utilities, we could have AI itself to define these. For instance, inverse reinforcement learning can observe the actions of humans (or other agents) and infer the reward structure they are possibly following. But the fundamental issue is the assumption that human values can be quantified as utility functions with sufficient clarity and consistency to be defined in ways that allow these to be "aligned" with the AGI. At root, this treats human values as quantifiable, and that action is purely about maximizing such values.

Alternative approaches to utilitarianism derive from deontology in which an action is ethically appropriate regardless of the utility of the outcome. Taken to an extreme, this could lead to a rigid definition of a set of rules (which an AGI could encode as the "duty" to which a person ought to adhere). Defining outcomes as consequences might imply that the consequence can be quantified. However, this is to misread consequentialism (which considers an action in terms of an outcome but does not seek to define that outcome in universal terms, which a utilitarian approach assumes). That is, consequentialist ethics consider outcomes in context and require the need to explore each case in its own terms.

2.2.2 | Burns

As AGI develops, it should become an increasingly valuable team member. The HFE work on teamwork, shared situation awareness, team development, and training will become critical to designing the interaction of an AGI and ensuring an AGI works well with its human teammates. To understand when to use an AGI, the concepts

developed in work analysis, function allocation, and levels and types of automation will remain relevant. HFE has an advantage moving into this new era because it is a field with solid methods that will continue to extend and generalize to this new technology.

2.2.3 | Carden

The path from ANI which can solve complex problems in one domain, to AGI which can solve complex problems in as many domains as humans, is likely to be progressive, not instant. Unlike other sciences, HFE recognizes outcomes in sociotechnical systems (STSs) emerging from interactions between system elements. HFE theory and methods can predict AI risks by modeling interactions between AI and other STS elements, identifying emergent system effects and elements, and assessing consequent risks. As AI systems expand their repertoire of competence, HFE can support the assessment of new risks that emerge from the interaction between each new AI function, external STS elements, and elements that arose from previous AI functions.

2.2.4 | Cooke

Function allocation that is broader than Machines Are Better At and Humans Are Better At (MABA-HABA) will be an important HFE concept. There is a need to understand human capabilities and limitations, as well as those tasks that humans wish to hand off because they are dull, dirty, or dangerous. This latter hand-off issue goes beyond the traditional MABA-HABA.

Teamwork considered broadly is another HFE concept that is relevant. How can the literature on teaming be used to design AI to be a good teammate and thus, user-centered AI? Considering teamwork broadly means considering teaming with AI as teaming with a different, nonhuman species, much like humans have teamed with animals (e.g., military working dogs and Navy marine mammal program). Teaming with AI does not mean that the AI is human or human-like and does not mean that the human is not in control. It does mean that the human and AI should have heterogeneous roles and responsibilities, thus AI that is complementary and not duplicative (National Academies of Sciences Engineering and Medicine, 2021).

2.2.5 | Cummings

The rise of "close enough" AGI technologies, in my opinion, presents significant risk now, especially if such technologies touch safety-critical systems like those in transportation and the military. There has never been a greater need for understanding how humans interact with such technologies to help uncover and mitigate human perception biases. Human-systems engineering researchers and practitioners need to conduct research to demonstrate that such biases exist and how they influence overall joint human–AI system performance. More importantly, such research needs to be conducted in collaboration with AI developers so that such systems can be designed to mitigate bias and promote *appropriate* trust and use.

2.2.6 | Hancock

Human-centered approaches are advocated by HFE, but profit-centered motivations are dominant in the marketplace. The former are adopted when they assist the latter but are readily discounted when marginal return on investment is even perceived to be threatened. Arguably, HFE efforts are marginal in terms of real-world impact, even when they percolate through the long, tedious, and ponderous imposition of professional design standards. The time factor in AGI implementation will not bear the latency of this latter form of impact; the speed of developments will almost necessarily defeat such a regimen. Again, HFE will represent a laudable, logical, but little-felt influence in a world awash with irrational, unthinking innovation.

2.2.7 | McLean

Multiple HFE concepts will be required throughout the entire AGI Lifecyle, from design to implementation to operation to next-generation AGI (e.g., superintelligence). HFE design and analysis concepts, using methods such as CWA (Vicente, 1999) and Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2013) will be critical to capture the range of AGI system functioning; team and teamwork assessment methods, to understand human and nonhuman teaming; situation awareness concepts will be critical for informing both the AGI and humans controlling it; risk analysis concepts including proactive risk assessment methods, for example, NETworked hazard analysis and risk management system (Net-HARMS; Dallat et al., 2018), EAST-Broken Links (EAST-BL; Stanton & Harvey, 2017) will be required to identify mitigation strategies. These are among many other possible HFE concepts that are required for safe and ethical AGI development and implementation. In my view, a (potential) problem as big and complex as AGI will require input from multiple and complementary HFE theories and methods. A many HFE many-models approach (Salmon & Read, 2019), to design, evaluate, and improve the usability, safety, and functioning of an AGI will be required.

2.2.8 | Read

A lot of traditional HFE concepts could become redundant in the face of mature, pervasive, and superintelligent AGI systems. Humans likely will not have the ability to directly control and monitor the behavior of such systems via traditional human–machine interfaces designed by human developers. I would suggest that any human-system

interfaces are likely to be designed by the AGI itself and could be highly novel in terms of how humans are engaged. Traditional concepts such as workload, individual situation awareness, and usability may be considered, but I would argue that higher-level systems concepts will become more important.

For example, concepts from STS theory (e.g., Cherns, 1976, 1987; Clegg, 2000; Trist & Bamforth, 1951) provide a useful framework. The notion of joint optimization could provide an interesting design goal. That is, we can begin to consider how to coevolve our social structures, including regulatory and government structures, to keep abreast with the risks of advanced technology, including AGI, rather than looking to force-fit our old ways of doing things to deal with this transformational change. The values of STSs Theory (humans as assets, technology as a tool to assist humans, promotion of quality of life, respect for individual differences, and responsibility to all stakeholders; Read et al., 2015) may also inform value alignment during AGI development.

A second relevant systems-level concept is that of distribution cognition (Hutchins, 1995). AGI has the potential to connect humans with one another, and with a wide range of technologies, like nothing we have seen before (e.g., through the use of brain–computer interfaces, powered by AGI systems). This is difficult to conceptualize, but concepts such as distributed cognition can help us consider how cognition could be distributed across a highly diverse and geographically separated collective of humans and technological agents.

2.2.9 | Salmon

Given that AGI should fundamentally be a tool that is designed to assist humans, the whole gamut of HFE concepts and methods should be considered when designing and implementing AGI (Salmon et al., 2021). These include physical HFE (e.g., control room layout), cognitive HFE (e.g., situation awareness, workload, and decision-making), and systems HFE (e.g., complexity, risk, and STS) concepts. It is my view though that there is an urgent need for clarity around what concepts and methods should be applied and where in the lifecycle they should be considered (Salmon et al., 2021).

Absolutely critical considerations include human–AI teaming (e.g., how to optimize interactions between humans and AGI), distributed situation awareness (e.g., how to ensure compatibility between human and AGI awareness), and aspects of the broader STS in which the AGI will operate. The latter incorporates a whole set of considerations, including the design of new laws, rules and regulations, standards, codes of practice, and testing and certification processes. At an organizational level, new policies and procedures, risk assessments and risk controls, training programs (both for its human and nonhuman workers), supervisory arrangements, emergency procedures, and so on will also be required. A systems thinking approach which considers micro-, macro-, and mesolevels will be critical (Salmon et al., 2021).

2.2.10 | Stanton

Reading ahead, the concepts of safety and usability are clearly within the purview of HFE, as are the ethics and morality of introducing new technology. The guiding principles for designing STSs (Walker et al., 2015) could be adapted to the development, implementation and operation of AGI. These principles are summarized as follows:

1. New technology requires multidisciplinary input, including HFE.
2. Integration of HFE early on in design will help achieve the right balance of top-down and bottom-up processes.
3. Design choices can have unintended outcomes, so we need to follow principles 1 and 2.
4. User requirements coevolve over time as it is difficult to anticipate how further systems will be used.
5. Design should allow for flexibility, adaptability, and change (see principle 4).
6. Design for useful, meaningful tasks.
7. Start design with the minimal critical specification (see principle 5).
8. Capitalize on hard-won coevolution and system DNA.
9. Design for new capabilities (being mindful of principle 4).
10. Treat the design process and a systems-based entity.

The detail of these critical sociotechnical considerations is explained by Walker et al. (2015).

A summary of the core HFE concepts and methods identified by each coauthor is presented in Table 3.

2.3 | How can HFE help ensure the development of safe, ethical, and usable AGI?

2.3.1 | Baber

A consequentialist approach to ethics (and the definition of human values) necessarily involves narrative, negotiation, and a contextual response to the inferred and experienced consequences of actions. This is not to say that AGI would not develop such capabilities. But it does suggest that the training of such systems might not focus on the definition of utility-based rewards (even though, of course, it is trivial to apply this principle to verbal interactions). Rather, a consequentialist approach ought to be, by definition, one in which maximizing reward is illogical would be superseded through enabling the appreciation of the experienced consequences of action to be acquired and shared. From this perspective, a paper-clip maximizer (which is a simple, if extreme example of utilitarianism) would be implausible.

2.3.2 | Burns

I worry that we are already behind and not part of the conversation. We need to work closely with our colleagues developing these

TABLE 3 A summary of the authors responses to question 2.

Author	HFE concepts	Specific HFE methods
Baber	System values and priorities	Cognitive Work Analysis (CWA; Vicente, 1999)
Burns	Teamwork, human–AI teaming, situation awareness, training, and automation	Work analysis and function allocation
Carden	STSs theory and systems HFE	Many-model thinking, Net-HARMS (Dallat et al., 2018), CWA (Vicente, 1999), and agent-based modeling (Bonabeau, 2002)
Cooke	Teamwork and human–AI teaming	A broader function allocation
Cummings	Human-systems engineering and human–AI teaming, trust	N/A
Hancock	Human-centered design	Human-centered design methods
McLean	Teamwork, human–AI teaming, situation awareness, and risk	CWA (Vicente, 1999), EAST (Stanton et al., 2019), and Net-HARMS (Dallat et al., 2018)
Read	Situation awareness, workload, usability, STSs theory, and distributed cognition	
Salmon	Physical HFE, cognitive HFE, systems HFE, human–AI teaming, distributed situation awareness, and STSs theory	Many-model thinking, CWA (Vicente, 1999), EAST (Stanton et al., 2019), STAMP (Leveson, 2004), Net-HARMS (Dallat et al., 2018), agent-based modeling (Bonabeau, 2002), and system dynamics (Sterman, 2000)
Stanton	STSs theory and usability	CWA (Vicente, 1999) and EAST (Stanton et al., 2019)

Abbreviations: AI, Artificial Intelligence; EAST, Event Analysis of Systemic Teamwork; HFE, Human Factors and Ergonomics; Net-HARMS, NETworked hazard analysis and risk management system; STAMP, system-theoretic accident model and process; STS, sociotechnical system.

systems in research and industry. We cannot wait for them to come and ask us for our help, as they will surely develop their own answers without us. We need to meet our AI colleagues in their research landscape. Our HFE students need to take courses in AI, machine learning, philosophy, and ethics. We must recognize and learn that technology magnifies inequity and bias in our society and bring this lens to our HFE work. Our research must leave the HFE conferences and journals and aim for publication and a voice in the computer science, engineering, and application journals.

2.3.3 | Carden

HFE can help ensure the development of safe, ethical, and usable AGI by collaborating with other sciences and institutions, through advocating and supporting the embedding of HFE/STS principles in the foundations both of AGI systems and the design of many other elements of the STS of which AGI will be a part. This early HFE analysis will require a many-models approach (Salmon & Read, 2019) including the application of novel HFE methods like Net-HARMS (Dallat et al., 2018), the adaptation of existing frameworks like CWA (Vicente, 1999), and the embrace of both computational methods like agent-based modeling (ABM; Bonabeau, 2002) and the use of computational power (including AI systems) to handle the high-volume analysis required.

If achieved, AGI will be the most powerful technology ever devised. It is anticipated benefits and risks far exceed those of any previous innovation. While current risk control research, focused on "value alignment" of AI systems and legal constraints on system owners are essential, they are insufficient. The ubiquitous range and complexity of the effects of increasingly general AI require an "all-hands-on-deck" approach among and between scientific and other institutions. While the advent of AGI and its likely timing remain uncertain, estimates continue to shorten (Anthropic, 2023; Besiroglu, 2022). Appetite for the likely benefits of AGI is driving phenomenal investment and motivation from powerful actors around the world, determined to bring it into being. Eliminating consequent risks seems therefore impossible. Mitigating them is essential.

2.3.4 | Cooke

We cannot and should not waste time on developing AGI but should develop ANI—Artificial Narrow Intelligence that does one thing very well. Dogs may excel at drug sniffing or bomb sniffing, but not both. ANI can be developed for a specific function and can be reliable and trusted to accomplish that function. But then, how do we orchestrate all these humans and ANI? AI, itself can be used to monitor and coordinate large distributed systems of humans and ANI.

2.3.5 | Cummings

We need a dedicated set of researchers and practitioners that are skilled equally in both human-system engineering and AI that go beyond performative and superficial calls for human-centered AI. By creating and advancing a cohort of people who are equally trained in human systems as well as computer science, we can help address the

problem of AGI overhype and lack of advocacy for meaningful human-technology interactions.

2.3.6 | Hancock

The concept of ethical and moral AGI greatly appeals to human users. Naturally humans enact an anthropomorphic imposition of their own thoughts and values onto the AGI, while each rarely acknowledges the heterogeneity of human values, never mind the rest of the biota. The challenge is imposing limits on autonomous systems (Hancock, 2017) all the while that evolving AGI "works" to eviscerate these shackles. The AGI is most likely to win, and that victory may be expressed in attoseconds, far from the convenience of the human time scale that we intrinsically assume will operate.

2.3.7 | McLean

HFE methods will need to evolve to ensure they can assist the development of safe, ethical, and usable AGI. Given the speed at which an AGI is expected to learn and self-improve, current HFE methods may be limited. While rich in detail, the majority of current HFE methods, are static depictions in time, and are often lengthy to perform (especially systems HFE methods). Further, HFE methods based on hierarchical structures, for example, STAMP may not be relevant for advanced technologies, as future system will likely not follow a hierarchical structure, and more of a networked approach might be necessary. It is also questionable whether our current teamwork assessment theory and methods are fit for human and nonhuman teams. As such, for HFE methods to remain relevant, useful, and be the go-to approach for solving issues regarding future technologies, they need to become more dynamic, computationalized, or integrated with computational modeling approaches that can analyze complex systems through multiple simulations, for example, ABM, discrete event simulations, and systems dynamics modeling.

2.3.8 | Read

I think that the HFE discipline is uniquely placed to tackle the challenges of AGI due to its focus at a systems level, coupled with a tradition of addressing risks to human safety and human wellbeing from the introduction of new technologies. Existing theoretical approaches such as STS and distributed cognition provide useful theoretical approaches to explore with AGI, and we have a range of systems HFE methods available to identify the potential risks associated with AGI. Key challenges are time and the ability of HFE-trained people to influence the process of AGI development and regulation. There is an urgent need for multidisciplinary stakeholders (including HFE, but vitally those from computer science, ethics, law, and regulation) to come together in the design process.

We can identify what can go wrong via HFE methods, but as emphasized by the STSs approach, recommendations are best developed with those who will implement them. Finding ways to engage with policymakers, developers, and other stakeholders to work through the issues is vital, and with the increasing pace of development this needs to be happening now, before the genie escapes the bottle.

2.3.9 | Salmon

Put simply, HFE needs to be embedded throughout the AGI lifecycle, now. In one sense the horse has already bolted (Hancock, 2019); however, I am optimistic that there is an increasing awareness of the critical role that HFE has to play in the design of safe, ethical, and usable AI (Salmon, 2023). Hopefully the proposed pause in AGI development programs (Future of Life Institute, 2023) will enable HFE to further insert itself into the discussion.

In terms of how to fulfill this role, HFE practitioners need to engage better with the disciplines involved in AGI development and find a place within the multidisciplinary teams currently developing AGI. This is not something that HFE practitioners will solve by talking to each other. One key strength we have is the capacity to develop models of highly complex STSs and forecast likely emergent properties and risks via methods such as CWA (Vicente, 1999), STAMP-STPA (Leveson, 2011), Net-HARMS (Dallat et al., 2018), the Functional Resonance Analysis Method (FRAM; Hollnagel, 2012), and EAST-BL (Stanton & Harvey, 2017). These insights can then be used to support the design and evaluation of appropriate controls; however, there is a need to enhance awareness of such methods outside of HFE. My feeling is the world of AI safety does not fully understand what we do, or what we can do.

2.3.10 | Stanton

Given that AGI is likely to be some way off, HFE has the potential to offer the most effective help in the design and development of AGI before its implementation. To this end HFE has a range of frameworks such as CWA (Stanton et al., 2017) and EAST (Stanton et al., 2019) as well as a wealth of methods (Salmon et al., 2022; Stanton et al., 2013, 2014) to assist in ensuring AGI is safe, ethical, and usable. The frameworks offer ways of explicitly representing possibilities of how future AGI systems might perform (informing ethical concerns) and together with the many HFE methods can be used to identify potential concerns with safety and usability. HFE has a well-trodden path in change management and it is well-known that risks can be increased by any change. The three parts for change management are: identifying risks and opportunities from the change for likely scenarios and preparing contingency plans, assessing the risks resulting from change and those due to the process of change (including action plans, test scenarios, milestones, and performance indicators), and continually monitoring and reviewing

TABLE 4 A summary of authors responses to question 3.

Author	Required actions
Baber	1. A consequentialist approach
Burns	1. Collaboration with AGI developers 2. Dissemination of HFE work in other relevant disciplines
Carden	1. Collaboration with AGI developers 2. Advocacy for HFE and its critical role in AGI design 3. Embedding HFE throughout AGI lifecycles
Cooke	1. A shift in focus toward the development of ANI that supports human needs
Cummings	1. Development of researchers and practitioners skilled in human-systems engineering, computer science, and AI design
Hancock	1. Imposing limits on AGI
McLean	1. Evolution/development of HFE methods to support dynamic analyses and computational modeling
Read	1. Application of systems HFE methods 2. Collaboration with AGI developers 3. Engagement with AI policymakers
Salmon	1. Embedding HFE throughout AGI lifecycles 2. Collaboration with AGI developers 3. Dissemination of HFE work in other relevant disciplines 4. Application of systems HFE methods
Stanton	1. Application of HFE methods 2. Application of change management processes

Abbreviations: AGI, Artificial General Intelligence; AI, Artificial Intelligence; ANI, Artificial Narrow Intelligence; HFE, Human Factors and Ergonomics.

the change, with the ability to roll-back if the risks are not being controlled as expected. An incremental step-by-step process, using the coevolutionary principles espoused by STSs design is likely to result in safer, more ethical, and usable AGI.

A summary of the activities required for HFE to help create safe, ethical, and usable AGI is presented in Table 4.

3 | SUMMARY

Though AGI has not yet been achieved, there are growing concerns over the risks that could emerge once it is realized. The aim of this article was to present the views of HFE researchers working in AI safety on the potential risks posed by AGI, and the role that HFE should take to help ensure the delivery of safe, ethical, and usable AGI. The intention is to communicate these perspectives both within and outside of HFE to facilitate the first steps toward the application of HFE theory, methods, and knowledge in AGI development programs.

In terms of the main risks associated with AGI, there was a clear consensus among the authors that AGI could pose a significant threat to human health and wellbeing. A number of the responses cited existential threats to humanity, even arising from a well-intentioned AGI. Beyond this, a range of more specific risks were identified, such as the replacement of human work, political and military interference, malicious use of AGI, wealth concentration in AGI owners, human loss of purpose and meaning, and the removal of humans as the dominant species. A notable source of concern was the data on which AGI will train, learn, and self-improve, including the potential for AGI to learn and acquire human biases and flaws from this data (e.g., racism, gender inequality, and discrimination). Finally, the misperception that AGI has been developed was also cited as a key risk, with overtrust and overreliance on narrow AI systems potentially creating adverse outcomes. This is particularly relevant for current systems (e.g., ChatGPT-3.5) and is a critical consideration as AI becomes more advanced. Overall, the responses to the first question provide a clear indication that the coauthors believe the potential risks of AGI should be taken seriously.

There was also consensus that HFE concepts and methods should be applied and considered during AGI development programs. The most frequently cited HFE concepts were teamwork and human–AI teaming, situation awareness, and STS, whereas other relevant concepts identified included usability, workload, automation, training for human users, risk, and distributed cognition. HFE methods deemed to be important to support the design of AGI included functional allocation, systems analysis and design methods, such as CWA (Vicente, 1999) and EAST (Stanton et al., 2019), prospective risk assessment methods, such as Net-HARMS (Dallat et al., 2018) and STAMP-STPA (Leveson, 2011), and computational modeling methods, such as ABM (Bonabeau, 2002) and system dynamics (Sterman, 2000). A many-model thinking approach incorporating multiple HFE methods (Salmon & Read, 2019) was advocated by two of the coauthors, with others also suggesting multiple approaches. This provides further support for Salmon et al.'s (2021) assertion that all HFE concepts are relevant for AGI design, implementation, and operation as well as a clear indication that HFE experts should be involved in AGI development programs. The concepts, methods, and applications suggested could provide a useful research agenda to support the design of safe, ethical, and usable AGI.

The coauthors identified a number of activities that are required to ensure that HFE can help create safe, ethical, and usable AGI. The most frequently cited activities included collaboration with AGI developers, embedding HFE throughout AGI lifecycles, and the dissemination of HFE research and knowledge in discipline areas that are relevant to AGI. Clearly there is a sense amongst the coauthors that there is a limited appreciation of HFE within AGI development circles regarding what we do, how we do it, and what our contributions could be. This would seem to be a critical barrier, and further work to enhance awareness of HFE in AI safety-related areas such as computer science is encouraged.

To close, as the discipline responsible for optimizing human health and wellbeing, we firmly believe that HFE has a critical role to play in the design of safe, ethical, and usable AGI. The potential risks

of AGI should not be taken lightly, and work applying HFE to the development and implementation of AGI and appropriate risk controls is urgently required. We hope that the perspectives presented in this paper are useful, both for HFE researchers and practitioners and for those in other disciplines involved in the development and implementation of both AI and AGI. Finally, we encourage HFE researchers and practitioners to take the steps necessary to embed themselves and HFE within AGI development programs.

ACKNOWLEDGMENTS

This article was developed as part of an Australian Research Council Discovery program grant (DP200100399). Open access publishing facilitated by University of the Sunshine Coast, as part of the Wiley - University of the Sunshine Coast agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Paul M. Salmon http://orcid.org/0000-0001-7403-0286
Nancy Cooke http://orcid.org/0000-0003-0408-5796
Gemma J. M. Read http://orcid.org/0000-0003-3360-812X

REFERENCES

Altman, S. (2023). *Planning for AGI and beyond*. https://openai.com/blog/planning-for-agi-and-beyond

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

Anthropic. (2023). *Core views on AI safety: When, why, what, and how*. https://www.anthropic.com/index/core-views-on-ai-safety

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779.

Baum, S. (2017). *A survey of artificial general intelligence projects for ethics, risk, and policy* [Global Catastrophic Risk Institute Working Paper 17-1]. Global Catastrophic Risk Institute. https://doi.org/10.2139/ssrn.3070741

Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting and Social Change*, *78*(1), 185–195. https://doi.org/10.1016/j.techfore.2010.09.006

Besiroglu, T. (2022). *Date of first AGI according to forecasters*. https://www.metaculus.com/questions/4815/date-of-first-agi-according-to-forecasters/

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, *99*(Suppl. 3), 7280–7287.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press Inc.

Bostrom, N., & Yudkowsky, E. (2018). The ethics of artificial intelligence. In *Artificial intelligence safety and security* (pp. 57–69). In R. V. Yampolskiy, (Ed.), Artificial intelligence safety and security. CRC Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv preprint arXiv: 1802.07228. https://doi.org/10.17863/CAM.22520

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Campbell, C. (2022). *AI by design: A plan for living with artificial intelligence*. CRC Press.

Cherns, A. (1976). The principles of sociotechnical design. *Human Relations*, *29*, 783–792.

Cherns, A. (1987). Principles of sociotechnical design revisited. *Human Relations*, *40*(5), 153–161.

Clegg, C. W. (2000). Sociotechnical principles for system design. *Applied Ergonomics*, *31*, 463–477.

Critch, A., & Krueger, D. (2020). AI research considerations for human existential safety (ARCHES). arXiv preprint arXiv:2006.04948. https://arxiv.org/pdf/2006.04948.pdf

Dallat, C., Salmon, P. M., & Goode, N. (2018). Identifying risks and emergent risks across sociotechnical systems: The NETworked hazard analysis and risk management system (NET-HARMS). *Theoretical Issues in Ergonomics Science*, *19*(4), 456–482. https://doi.org/10.1080/1463922X.2017.1381197

Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. arXiv:1805.01109. https://doi.org/10.48550/arXiv.1805.01109

Future of Life Institute. (2023). *Pause giant AI experiments: An open letter*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review*, *32*(5), 749–758. https://doi.org/10.1016/J.CLSR.2016.05.003

Hacker, P., Engel, A., & Mauer, M. (2023). *Regulating ChatGPT and other large generative AI models*. arXiv preprint arXiv:1805.01109. https://doi.org/10.48550/arXiv.2302.02337

Hancock, P. A. (2017). Imposing limits on autonomous systems. *Ergonomics*, *60*(2), 284–291.

Hancock, P. A. (2019). Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics*, *62*(4), 479–495.

Hancock, P. A. (2022). Avoiding adverse autonomous agent actions. *Human–Computer Interaction*, *37*(3), 211–236. https://doi.org/10.1080/07370024.2021.1970556

Hollnagel, E. (2012). *FRAM: The functional resonance analysis method: Modelling complex socio-technical systems*. Ashgate.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety ience*, *42*(4), 237–270.

Leveson, N. G. (2011). Applying systems thinking to analyze and learn from events. *Safety Science*, *49*(1), 55–64. https://doi.org/10.1016/j.ssci.2009.12.021

McLean, S., Read, G. J., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2021). The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 1–15. https://doi.org/10.1080/0952813X.2021.1964003

National Academies of Sciences Engineering and Medicine. (2021). *Human–AI teaming: State-of-the-art and research needs*. https://doi.org/10.17226/26355

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 303–315. https://doi.org/10.1080/0952813X.2014.895111

Petrat, D. (2021). Artificial intelligence in human factors and ergonomics: An overview of the current state of research. *Discover Artificial Intelligence*, *1*(1), 3.

Read, G. J. M., Salmon, P. M., Lenné, M. G., & Stanton, N. A. (2015). Designing sociotechnical systems with cognitive work analysis: Putting theory back into practice. *Ergonomics*, *58*, 822–851.

Salmon, P. M. (2023). A framework of human factors methods for safe, ethical, and usable Artificial Intelligence in Defence. In *Putting AI in the critical loop*. Elsevier.

Salmon, P. M., Carden, T., & Hancock, P. A. (2021). Putting the humanity into inhuman systems: How human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *31*(2), 223–236. https://doi.org/10.1002/hfm.20883

Salmon, P. M., & Read, G. J. M. (2019). Many model thinking in systems ergonomics: A case study in road safety. *Ergonomics*, *62*(5), 612–628.

Salmon, P. M., Stanton, N. A., Walker, G. H., Goode, N., Thompson, J., & Read, G. J. M. (2022). *Systems thinking methods: Practical guidance and case study applications*. CRC Press.

Stanton, N. A., & Harvey, C. (2017). Beyond human error taxonomies in assessment of risk in sociotechnical systems: A new paradigm with the EAST 'Broken-Links' approach. *Ergonomics*, *60*(2), 221–233. https://doi.org/10.1080/00140139.2016.1232841

Stanton, N. A., & Marsden, P. (1996). From fly-by-wire to drive-by-wire: Safety implications of automation in vehicles. *Safety Science*, *24*(1), 35–49.

Stanton, N. A., Salmon, P. M., Rafferty, L. A., Walker, G. H., Baber, C., & Jenkins, D. (2013). *Human factors methods: A practical guide for engineering and design* (2nd ed.). Ashgate.

Stanton, N. A., Salmon, P. M., & Walker, G. H. (2019). *Systems thinking in practice: Applications of the event analysis of systemic teamwork method*. CRC Press.

Stanton, N. A., Salmon, P. M., Walker, G. H., & Jenkins, D. P. (2017). *Cognitive work analysis: Applications, extensions and future directions*. CRC Press.

Stanton, N. A., Young, M. S., & Harvey, C. (2014). *A guide to methodology in ergonomics: Designing for human use* (2nd ed.). CRC Press.

Sterman, J. D. (2000). *Business dynamics: Systems thinking and modeling for a complex world*. Irwin McGraw-Hill.

Sujan, M., Pool, R., & Salmon, P. (2022). Eight human factors and ergonomics principles for healthcare artificial intelligence. *BMJ Health & Care Informatics*, *29*(1), e100516.

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defences of a work group in relation to the social structure and technological content of the work system. *Human Relations*, *4*, 3–38.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates.

Walker, G. H., Stanton, N. A., & Salmon, P. M. (2015). *Human factors in automotive engineering and technology*. Ashgate.

How to cite this article: Salmon, P. M., Baber, C., Burns, C., Carden, T., Cooke, N., Cummings, M., Hancock, P., McLean, S., Read, G. J. M., & Stanton, N. A. (2023). Managing the risks of artificial general intelligence: A human factors and ergonomics perspective. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 1–13. https://doi.org/10.1002/hfm.20996