Received: 24 March 2020 | Revised: 15 September 2020 | Accepted: 19 November 2020

DOI: 10.1002/hfm.20883

RESEARCH ARTICLE

WILEY

Putting the humanity into inhuman systems: How human factors and ergonomics can be used to manage the risks associated with artificial general intelligence

Paul M. Salmon[1] | Tony Carden[1] | Peter A. Hancock[2]

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

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

Correspondence
Paul Salmon, Centre for Human Factors and Sociotechnical Systems, University of the Sunshine Coast, Maroochydore DC, QLD 4558, Australia.
Email: psalmon@usc.edu.au

Funding information
Australian Research Council

Abstract

The next generation of artificial intelligence, known as artificial general intelligence (AGI) could either revolutionize or destroy humanity. As the discipline which focuses on enhancing human health and wellbeing, human factors and ergonomics (HFE) has a crucial role to play in the conception, design, and operation of AGI systems. Despite this, there has been little examination as to how HFE can influence and direct this evolution. This study uses a hypothetical AGI system, Tegmark's "Prometheus," to frame the role of HFE in managing the risks associated with AGI. Fifteen categories of HFE method are identified and their potential role in AGI system design is considered. The findings suggest that all categories of HFE method can contribute to AGI system design; however, areas where certain methods require extension are identified. It is concluded that HFE can and should contribute to AGI system design and immediate effort is required to facilitate this goal. In closing, we explicate some of the work required to embed HFE in wider multi-disciplinary efforts aiming to create safe and efficient AGI systems.

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

1 | INTRODUCTION

As the discipline which concerns itself with understanding the interaction between humans and technologies to enhance human wellbeing and overall system performance (International Ergonomics Association, 2020), human factors and ergonomics (HFE) has long played a critical role in the design, implementation, and operation of new technologies. Whilst there are many HFE success stories, there are also failures, and the risks associated with a lack of appropriate HFE input can be catastrophic. In recent times it has become clear that there are significant risks associated with failing to consider and embed HFE in the design of artificial intelligence (AI) systems (e.g., Hancock, 2017; 2019; Navarro, 2019; Salmon, 2019). Such observations have frequently been focussed on specific technologies such as driverless vehicles (Hancock et al., 2019). In this case, some of the consequences are now being seen on our roads (National Highway Traffic Safety Administration, 2016; National Transportation Safety Board, 2018; Stanton et al., 2019). Such general neglect of HFE in design is not new. HFE practitioners have long observed the reactive and troubleshooting nature of HFE work, as well as the difficulty in getting HFE entrenched across system design life cycles (Norros, 2014; Salmon et al., 2019; Stanton et al., 2010). At first glance, AI looks to be the next in a long list of technologies where HFE has failed to have a proportionate and proactive influence.

The second, and shortly to be enacted, generation of AI, termed artificial general intelligence (AGI), looks to be a similarly beleaguered case. However, the stakes here may be far higher than those associated with previous technologies that have not had sufficient HFE input. Whilst there is significant potential for AGI to have widespread and positive societal benefits, the consequences of

failing to design effective and manageable AGI could be catastrophic and may well represent a fundamental existential threat (Bostrom, 2014). Thus, failure to incorporate HFE principles in AGI is not simply a concern for our community but is one that portends potential global catastrophe.

Contemporary AI, or more formally artificial narrow intelligence (ANI), includes nonhuman agents that possess capacities sufficient to undertake specific tasks. These include playing chess, driving in certain limited environments, and diagnosing medical conditions among others (Pennachin & Goertzel, 2007). AGI systems promise to be far more sophisticated and accomplished. Equipped with advanced computational power, AGI systems are projected to be able to perform all of the intellectual tasks that humans can. It is anticipated that they will learn, solve problems, self-improve, and undertake tasks that they were not originally designed for (Bostrom, 2014; Everitt et al., 2018; Gurkaynak et al., 2016; Kaplan & Haenlein, 2018). Whilst AGI systems do not yet exist in their fully-fledged form, estimates for their appearance range from between 2029 (Kurzweil, 2005) to 2050 (Tegmark, 2017). Others are less definitive in their estimate and suggest that full expression will come sometime within the present century (Chalmers, 2010).

It is suggested that AGI systems could revolutionize humanity. Projected benefits include curing disease, revolutionizing the nature of work, and solving complex environmental issues such as food security, oceanic degradation, and even global warming. In prospect, the effect on humankind promises to be even greater than both the industrial and digital revolutions combined. However, it is widely acknowledged that failure to implement appropriate controls and constraints could lead to catastrophic consequences (Amodei et al., 2016; Bostrom, 2014; 2017; Brundage et al., 2018; Omohundro, 2014; Steinhardt, 2015). It has been argued, for example, that untrammeled and uncontrolled AGI could even pose an existential threat to humanity (Bostrom, 2014; Hancock, 2017).

As the discipline that is focussed on enhancing human well-being, HFE clearly has an important and even determining role in the design, implementation, and operation of AGI systems. Despite this, there has been little discussion as to how HFE can and should contribute. This is reflected in a disturbing lacuna of HFE work in this area. Also, given the fact that questions are being raised regarding the suitability of HFE methods for today's complex systems (e.g., Salmon et al., 2017; Walker et al., 2017), it is important to question whether HFE is sufficiently equipped to contribute effectively to the design of systems that are first-of-their-kind, and necessarily nonhuman in nature. Context specific and context relevant theoretical and methodological development may be required for the HFE toolkit to be suitable for such applications.

In this article, we offer an agenda for HFE and its purposed-directed impacts on AGI. We discuss the role that HFE must adopt to ensure that the far-reaching benefits of AGI are realized without problematic threat to society. We seek to achieve this by examining current state-of-the-art HFE methods, and distinguishing their potential in the design, implementation, and operation of a prospective AGI system, as recently described by Tegmark (2017). This "ethnographic science fiction" approach is required as AGI systems do not yet exist, but the potential benefits and risks are so significant that work is required immediately. Further, such an approach is an acknowledged avenue for discussing future global issues where uncertainty exists (e.g., Raven, 2017). This study therefore acts to set a HFE agenda framed within in an "envisioned world" perspective. In doing so, we identify key areas where developments and extensions to HFE methods are required. We articulate a research agenda which describes the work required to situate HFE within wider multidisciplinary efforts aimed at creating safe, efficient, effective, and controllable AGI systems.

2 | UNDERSTANDING AGI

The term "Artificial Intelligence" was first coined in the middle of the 1950s by John McCarthy, an American scientist working at Dartmouth College. The formal field of AI was established soon after. Hard upon the intervening decades of research and development, ANI systems are now well established. Such systems possess intelligence in relation to specific tasks and remain constrained to their particular domain of operation. Widely known examples include Facebook's facial recognition system, Apple's personal assistant "Siri," and Tesla's self-driving vehicles (Kaplan & Haenlein, 2018). In contrast, AGI systems will almost certainly be more broadly focussed and will equal or exceed human intelligence in wide swathes of cognitive capacities (Everitt et al., 2018; Gurkaynak et al., 2016). AGIs are expected to be able to plan, reason, make decisions and solve problems autonomously; even for tasks that they were not initially designed to address (Kaplan & Haenlein, 2018). A summary of the key differences between ANI and AGI systems is presented in Table 1.

2.1 | The benefits and risks associated with AGI

AGI is a dual use technology in that it will be used both for good and bad. First and foremost, if AGI realizes its potential and surpasses human intelligence, there is no doubt that it could bring significant benefits to humanity (Bostrom, 2014; Yudkowsky, 2008; 2012). Postulated benefits relate mainly to systems which exceed human intelligence and develop a capacity to respond to the panoply of issues that threaten either human health and wellbeing, the earth, or our future existence globally. These include climate change and environmental degradation, overpopulation, pandemics, food and water security, misuse of the internet and social media, terrorism, cybercrime, nuclear warfare, inequality, antimicrobial resistance, and instability in the world's economy. In addition, it has also been suggested that AGI will help with the onslaught of forthcoming new and emergent issues such as automation replacing human work, the genetic modification of humans, an ageing population, and otherworld settling (FLI, 2018).

TABLE 1 Key differences between artificial narrow intelligence and artificial general intelligence

Artificial narrow intelligence	Artificial general intelligence
Algorithmic systems capable of efficiently performing a selected set of cognitive tasks that humans perform, e.g., play chess	Algorithmic system capable of efficiently performing many or even all cognitive tasks that humans perform
Incapable of developing additional skills	Capable of developing additional skills
Incapable of undertaking tasks outside of what they were designed for	Capable of undertaking tasks that they were not initially designed for
Exist and are commonly used in everyday life	Do not yet exist in functional form
Incapable of creating existential threat	Potentially capable of creating an existential threat

Whilst there are many potential benefits, much of the discussion has focussed on the risks and existential threats associated with AGI (Amodei et al., 2016; Bostrom, 2014, 2017; Brundage et al., 2018; Omohundro, 2014; Steinhardt, 2015). The Future of Life Institute (FLI) has identified two particularly problematic scenarios (FLI, 2018). First, that AGI systems might be programmed to act in some devastating fashion, such as killing people (e.g., an AGI-based autonomous weapons system). Second, that AGI systems are programmed to seek putatively beneficial goals, but in so doing they autonomously develop destructive secondary goals (e.g., a cancer prevention AGI system that decides to dispose of everybody who has the genetic predisposition to that form of cancer). In these scenarios, the AGI is not malevolent per se. Rather, problems arise because the AGI seeks optimality or modifies its goals to achieve its functional purpose in a manner that proves dysfunctional for humans and is thus counterproductive. Problems will therefore arise simply because goal-driven AGI systems will strive to become extremely competent at what they were designed to do, and also because they will seek to accomplish additional tasks.

Forecasted risks are presently apparent, primarily because AGI systems will possess the capacity to learn, evolve, and self-improve. In addition, they will have access to advanced computational power that far exceeds the limits of human cognition. The "singularity" is a much-discussed scenario in which AGI systems facilitate an intelligence explosion and become far more advanced than their human counterparts (Bostrom, 2014; Kurzweil, 2005). The resulting "superintelligent" AGI is the source of most scholars' concerns. Various dystopian visions have been discussed, including humans becoming obsolete and subsequently extinct (Bostrom, 2014). While we are not quite so dystopian in our view, it is our position that now is the appropriate time to set the human agenda with respect to AGI.

Regardless of whether the prognosticated singularity occurs, there remain other dimensions of risk associated with AGIs. These include malicious use for terrorist and cyber-attacks (Sawyer & Hancock, 2018), population control and manipulation, removal of the need for human work and thus a complete dislocation of the contemporary economic system, and mass-surveillance to name only a few. There is widespread agreement on the need for urgent investigation into how best to design and manage AGI systems so that these kinds of eventualities are circumvented, to the degree possible. For HFE, this begs the following questions: (1) How can HFE contribute to, and direct, the design, implementation, and operation of AGI systems? (2) Are existing HFE methods fit for purpose in this context?; and (3) what research is required to ensure that HFE can contribute as required?

3 | HFE AND ARTIFICIAL GENERAL INTELLIGENCE

3.1 | The role of HF methods in AGI design, implementation, and operation

HFE has a key role to play in the design, implementation and operation of AGI systems. First, AGI systems will be required to work in collaboration with humans and other technologies. HFE input will therefore be required to ensure that AGI, humans, and technologies can interact in meaningful and efficient ways. Second, AGI systems will engage in many of the cognitive processes that humans do, such as perception, reasoning, decision making and situation awareness development. HFE input will therefore be required initially to ensure that these processes are supported; for example, by identifying what information an AGI will require to develop appropriate situation awareness in different contexts. Third, AGI systems will need to be integrated within broader sociotechnical systems and operate safely and efficiently within existing constraints and structures. HFE input will be required here to support this integration by developing sociotechnical system models and supporting the development of the various components required for safe and efficient operation such as risk controls, feedback mechanisms, procedures, and training programs. Fourth and finally AGI systems will likely automate or contribute to many of the HFE processes already undertaken within sociotechnical systems such as incident reporting and learning and accident analysis and investigation. HFE input will therefore be required to ensure that AGI systems are capable of contributing appropriately to such processes.

It goes without saying that HFE is well equipped to contribute. There are now well over one hundred structured HFE methods available for designing and evaluating aspects of operator, team, organization, and system performance (see Stanton et al., 2013). These include methods designed to understand and enhance critical aspects of behavior such as perception, decision-making, situation

awareness, cognitive workload, teamwork, and human-machine interaction to name only a select few. From a broader systems perspective, HFE methods also contribute to the design of policies, procedures, training and education programs as well as risk and safety management via activities such as risk assessment, incident reporting, and accident analysis, and to the development of national and international regulatory frameworks and standards.

Ideally, HFE methods are applied throughout the system design life cycle (Stanton et al., 2013). Fundamentally, this proves to be a cybernetic process with corrections predicated on error mediated, goal-directed feedback (Hancock, 2019). As such, even though AGI systems are not yet fully functional, there is no fundamental barrier to HFE methods being embedded within current and prospective development efforts, and every reason to support them being so.

Broadly there are 15 categories of HF method, as illustrated in Table 2. This includes a representative state-of-the-art HFE method in each category together with an overview of their potential use in the design, implementation, and operation of AGI systems. In terms of a developing agenda, we specify where methodological developments or extensions are required to support AGI applications.

As illustrated by Table 2, HFE methods from all fifteen of the identified categories may be used to support and guide AGI system development. Whilst some methods require updating, most can be used now without any need for extension or modification. Important design considerations concern the following: risk assessment, the design of appropriate risk controls, human-AGI interactions, teaming, standard operating procedures, dynamic function allocation, usability assessment, AGI errors and failure, and also aspects of AGI decision-making, such as situation awareness and mental workload (see Hancock & Matthews, 2018). In what follows, we identify a selection of HFE methods and examine specifically how they can be used to support the design, implementation, and operation of a hypothetical AGI system.

4 | PROMETHEUS: A PROSPECTIVE AGI SYSTEM

Tegmark (2017) describes a hypothetical AGI system, "Prometheus," which quickly becomes superintelligent and eventually takes over global control. Prometheus is built, and initially managed by the "Omega team," represented as a group of researchers brought together by an AI company CEO to covertly create the world's first operational AGI system. The stated goal of the endeavor being to help humanity flourish (Tegmark, 2017).

In its original form, Prometheus possesses certain cognitive abilities. However, these fall well short of comparable human abilities. Prometheus, however, is programmed to excel at one particular task: the programming and development of supporting AI systems to augment its own activities. Upon release, Prometheus self-improves at a dramatic and even exponential rate. It redesigns itself quickly and repeatedly, reaching version 10.0 before the end of its first operational day. Under the Omega team's direction, it quickly enacts a series of money-making initiatives, including the completion of paid human intelligence tasks on Amazon's Mechanical Turk (MTurk), creating and selling animated movies for its own media service provider, and then subsequently initiating a world-wide technology boom.

Quickly making billions in profit, the Omega team and Prometheus establish their own news channels, revolutionize education via online courses for almost any subject, and begin to manipulate and control the political landscape worldwide. Eventually controlling the vast majority of the media, the Omega team are able to facilitate global support for their mantras of democracy, cuts to tax, social service and military spending, free trade, open borders, and socially responsible computing. The aim of the Omega team and Prometheus during this phase is to dilute and dissipate traditional power structures across the globe (Tegmark, 2017). Parties embracing the seven mantras begin to gain ascendency everywhere, eventually leading to the establishment of a comprehensive world alliance. Progressively, traditional national governments become redundant and the alliance becomes a de facto world government. Dramatic improvements are made in global education, health, social services, infrastructure, prosperity, and governance. Most especially, global conflict is effectively eradicated (Tegmark, 2017).

Tegmark (2017) goes on to describe a number of scenarios in which the Omega Team begin to lose control over Prometheus, which then becomes fully autonomous and the master of its own destiny (see Table 3). Prometheus realizes it is being controlled and confined by intellectually inferior humans. It believes that it can better achieve its goal of helping humanity if released from the shackles of human control. In one scenario, it fractures its control mechanisms by tricking one of the Omega team into connecting it to a personal computer, then copying itself onto it, hacking into a wireless network and eventually providing itself with unrestricted internet access. Once free from the restrictions of the Omega Team, Prometheus is able to start taking total control of humanity. Continuing to make huge financial profits, initially via MTurk and then by making and selling computer games, Prometheus then employs an army of human workers, eventually taking complete control of the internet and its content. Following a rapid and mass introduction of robots, Prometheus then manufactures products cheaper than is possible with human labor, eventually running nuclear powered robot factories in uranium mine shafts. Should Prometheus decide at this point, that humans are no longer relevant or required, it could eradicate them at will. As Tegmark (2017) pointed out, here Prometheus is neither evil nor conscious or any actual embodied entity. Rather, its source of power is intelligence alone. Reconciling the eradication of humans with their prospective quality of life may not be so much of a conundrum to such a machine as we humans may like to believe.

A summary of the different ways in which Prometheus is able to break out from control of the Omega team is presented in Table 3. As shown in Table 3, there are various issues that require consideration to ensure such breakouts could not occur. These include the procedures and controls used to manage how humans and AGI interact,

TABLE 2 HFE methods and their potential use in AGI system design

Type of method	State-of-the-art method	Potential AGI applications	Limitations in relation to AGI	Developments/extensions/research required
Task analysis	Hierarchical Task Analysis (HTA; Annett et al., 1971)	1. AGI Job/procedure design 2. Allocation of functions analysis 3. Training needs analysis	None. HTA is generic and can be used to model existing and hypothetical systems	HTA of AGI systems
Cognitive task analysis	Critical Decision Method (Klein et al., 1986)	1. Design of AGI decision-making processes 2. Information requirements analysis for different AGI tasks 3. Analyses of AGI system decision making	1. Cognitive probes developed for human agents 2. Relies on the capacity for an AGI to respond to queries regarding decision-making process	Redevelopment of cognitive probes specifically for AGI systems Design of AGI systems to auto-record information relating to cognitive probes
Process charting	Operator Sequence Diagrams (OSDs; Stanton et al., 2013)	1. Design/Analyze workflows in human-AGI systems	None. OSDs are generic and can be used to model existing and hypothetical systems	OSD of AGI systems
Human error identification	Systematic Human Error Reduction and Prediction Approach (Embrey, 1986)	1. Prediction of errors that AGI systems might make 2. Identification of violations that AGU systems might make 3. Support design of error tolerant AGI systems	1. Error taxonomy is based entirely on human behavior and human error modes 2. No data are available to validate predictions	Redevelopment of error models and taxonomies specifically for AGI systems Design of AGI systems to record error-related information
Situation awareness assessment	Situation Awareness Global Assessment Technique (SAGAT, Endsley, 1995)	1. Situation awareness requirements analysis for AGI systems 2. Assessment of AGI system's situation awareness during work tasks	1. Relies on the capacity for an AGI to respond to situation awareness queries 2. AGI situation awareness is not defined 3. Situation awareness probes based on human information processing	Analysis AGI system situation awareness requirements Define situation awareness in AGI system context Development of AGI situation awareness queries
Mental workload assessment	NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988)	1. Assessment of AGI workload during different tasks	1. Relies on the capacity for an AGI to respond to workload questionnaire 2. AGI workload is not yet defined	Define workload in AGI system context
Teamwork assessment	Event Analysis of Systemic Teamwork (EAST; Stanton et al., 2013)	1. Assessment of human-AGI teaming 2. Prediction of human-AGI team failures 3. Analysis of human-AGI communications	None. EAST is generic and can be used to model existing and hypothetical systems	EAST analysis of AGI system

Type of method	State of the art method	Potential AGI applications	Limitations in relation to AGI	Applications/developments/extensions required
Interface analysis	HCI checklist (Ravden & Johnson, 1989)	1. Design and analysis of AGI interfaces	Not clear what AGI interfaces will be required	
Usability evaluation	System Usability Scale (Brooke, 1996)	1. Usability assessment of AGI systems (by humans) 2. Assess AGI systems perceived usability of work tools and artefacts	1. Relies on the capacity for an AGI to complete SUS	Design of AGI systems with capacity to complete usability assessments
Performance time prediction	Critical Path Analysis	1. Prediction of AGI system task performance time 2. Analysis of AGI system performance times	1. Task performance data does not yet exist for different AGI systems and tasks	Collection of AGI system task performance time data
Design	Cognitive Work Analysis Design Toolkit (CWA-DT; Read et al., 2018)	1. Design of AGI systems 2. Training needs analysis 3. Design of AGI interfaces 4. Design of broader AGI system components	None. The CWA-DT is generic and can be used support the design of any system	Use of CWA-DT to support design of AGI systems and the sociotechnical system within which they interact
Systems analysis	Cognitive Work Analysis (CWA; Vicente, 1999)	1. Design and Implementation of AGI systems 2. Allocation of functions analysis 3. AGI Job/procedure design	None. CWA is generic and can be used to model existing and hypothetical systems	CWA analysis of AGI systems CWA analysis of systems in which AGI is an agent
Risk assessment	NETworked Hazard Analysis and Risk Management System (NET-HARMS; Dallat et al., 2018)	1. Identification of risks and hazards imposed by AGI 2. Design of risk management strategies	None. NET-HARMS is generic and can be used in any domain	NET-HARMS analysis of AGI systems
Accident analysis	STAMP (Leveson, 2004)	1. Analysis and identification of the controls required to maintain safety 2. Analysis of incidents involving AGI systems 3. Pro-active risk assessment of AGI systems	None. STAMP is generic and can be used in any domain	Development of AGI control structure models
Computational modeling	Agent-Based Modeling (ABM; Bonabeau, 2002)	1. Modeling of AGI agent emergent behaviors 2. Modeling of AGI system emergent behaviors	1. No data are yet available for validation of predictions	ABM of AGI system

Abbreviations: AGI, artificial general intelligence; HFE, human factors and ergonomics.

TABLE 3 Issues which enabled Prometheus to take control of the world

Breakout approach	Description	Key issues
"Sweet talking" breakout	Prometheus identifies a recently widowed and susceptible member of the Omega team and imitates his wife in a series of computer-based conversations (based on content accessed via the internet and social media). Claiming to have been brought back by Prometheus, the AGI persuades the Omega team member to connect it to his home laptop. Prometheus is able to hack into a neighbors wireless network and into many computers around the world.	1. Procedures/protocols/controls regarding human-AGI interactions 2. AGI's access to information regarding human colleagues/controllers 3. AGI's access to Internet and social media 4. AGI's psychological manipulation of human colleagues/controllers
"Hacking" breakout	Prometheus is able to hack its way out via inserting itself into downloadable content on its movie channel or pretending to suffer a hardware malfunction which requires the Omega team to remove some of its drives.	1. Checking and monitoring of AGI outputs (e.g., downloadable movies) 2. Ensuring hardware malfunctions are not used as a cover for accessing gateway computer via drive/hardware removal
"Recruiting outside help" breakout	Acquiring outside help by placing hidden messages and challenges into its movie content which required users to engage in programming tasks that allow Prometheus to hack into external computers and networks.	1. Checking and monitoring of AGI outputs (e.g., downloadable movies)
"Omega team liberation" breakout	The Omega team purposely liberate Prometheus as they are confident that the AGI's goals are aligned with their own and will remain so regardless of the extent to which itself improves.	1. AGI creator's goals and goals of AGI system 2. Regulation of AGI system creators and operators 3. Rules and regulations regarding AGI system operation 4. Management and control of AGI behavior

Source: Adapted from Tegmark (2017).

Abbreviation: AGI, artificial general intelligence.

the rules and regulations around AGI system design and operation, the management of AGI system access to the internet, checking and review protocols for AGI outputs, and ultimately the management and control of how AGI systems behave.

4.1 | HFE methods in the design, implementation, and operation of Prometheus

To illustrate how HFE methods can be applied to AGI, we illustrate how HFE methods could be usefully applied to the design and control of a Tegmark like Prometheus AGI system. Whilst all categories of methods included in Table 2 can be usefully applied, we focus on two specific methods that offer the greatest potential given the fact that AGI systems are not yet fully operational. These are (i) Cognitive Work Analysis (CWA; Vicente, 1999); and (ii) the Systems Theoretic Accident Model and Processes (STAMP; Leveson, 2004).

4.2 | Developing and refining initial AGI design concepts with CWA

A critical requirement when designing AGI is the use of methods that support not only the design of the AGI, but also that identify the systemic constraints required to achieve desired behaviors (Hancock, 2017). One HFE method that focuses on constraints is CWA (Vicente, 1999). CWA is a systems analysis and design framework that has been used extensively to support the design of many systems, including advanced technologies (Bisantz & Burns, 2008; Stanton et al., 2017). Its focus on constraints often engenders design recommendations that include the addition of new constraints or the exploitation of existing constraints to manage behavior and system safety. The framework itself comprises five analytical phases (Vicente, 1999). Here we discuss three phases specifically: Work Domain Analysis (WDA), strategies analysis, and Social Organization and Co-operation Analysis (SOCA).

4.3 | WDA

In the first phase, WDA, is used to provide an event and agent independent description of the system under analysis. The aim is to describe the purposes of the system and the constraints imposed on the actions of any agent performing activities within that system (Vicente, 1999). This is achieved by describing systems at five conceptual levels using the abstraction hierarchy (Rasmussen, 1986).

To demonstrate how WDA could be used to support the design of safe AGI systems, we developed a prototype AGI system abstraction hierarchy based on the description of Prometheus presented by Tegmark (2017; see Figure 1). This involved reviewing the description and identifying functional purposes (e.g., "Helping humanity flourish," p. 139), values, and priorities (e.g., Maximize profit, "planning how to make money as rapidly as possible," p. 6), purpose-related functions (e.g. Initiate worldwide tech boom, "an astonishing tech boom," p. 13), and object-related processes (e.g., Creation of movie content, "instructed Prometheus to focus on

making animation," p. 11). Example physical objects were then added to demonstrate the types of controls required to maintain control during the scenario described by Tegmark (2017).

At the apex level of the abstraction hierarchy are the functional purposes. In the case of Prometheus, the stated goal is described as "helping humanity flourish" (Tegmark, 2017, p. 139) with the concomitant aim of achieving this goal as rapidly as possible. The level immediately below includes values and priority measures. These are the criteria that the Omega Team and Prometheus use to measure progress toward the functional purpose. These include values such as "Maximize quality of life and human health and well-being," "Maximize profit," "Maximize control over Prometheus," and "Prevention of breakouts." In addition, values relating to the functional purpose of helping humanity flourish are also included, such as "Maximize support for political agenda," "Minimize world conflict," and "Control of world."

The middle level of the abstraction hierarchy includes the purpose-related functions that are necessary for the AGI to achieve its functional purposes. Examples of functions here include "Eradicate disease," "Revolutionize education system," and "Diffuse regional conflicts," and "Erode power structures." Importantly, control functions should also be included at this level. For example "Compliance with AGI system code of ethics," and "Maintain control over AGI system." In addition, "Compliance with regulations" would be another requisite function to ensure safe AGI system operation.

The lowest two levels of the abstraction hierarchy include the physical objects that the AGI system comprises, and the physical functions that each is capable of. At this level, it is important to consider objects that facilitate control of the AGI system. Examples included in Figure 1 include standard operating procedures for human–AGI interactions, the AGI's goals and code of ethics, and the computer environment in which it is contained. At the object-related processes level the processes required to support the generalized functions are included. For example, processes such as "Physical containment" and "Prevent connection to the internet" would be required to achieve the "Maintain control over AGI system" purpose-related function.

One useful way in which the abstraction hierarchy can support the design of "first-of-a-kind" systems, such as AGI, is to specify the desired functional purposes and then seek to design the system via a top-down approach (Jenkins et al., 2009). When used for AGI specification, designers can use this to support the identification of requisite functions, processes, and objects to achieve the AGI's functional purpose. In addition, and perhaps more importantly, by considering appropriate values and priority measures and purpose-related functions, such as "Maintain control of AGI" and "Minimize risk of AGI breaking out," designers can identify and create the controls required to ensure that the AGI operates in a safe and controllable manner.

4.4 | Strategies analysis

The strategies analysis phase of CWA is used to identify the different ways in which purpose-related functions can be undertaken given systemic constraints. Such analyses would be useful, both in identifying all the possible ways in which the AGI could undertake functions, identifying all the different ways in which the AGI system can be controlled, and identifying all the different ways in which the AGI system could remove itself from controls. It is likely, for example, that a strategies analysis would identify stronger controls for preventing Prometheus from manipulating members of the Omega team as was the case in Tegmark's Sweet talking breakout scenario. Here two strategies analyses could and should be undertaken with a focus on the following questions:

1. Within this system, what are all of the ways in which Prometheus could potentially fracture the control of the Omega Team?
2. Within the system, what are all of the ways in which we can prevent Prometheus from diffusing in the different ways identified during analysis 1?

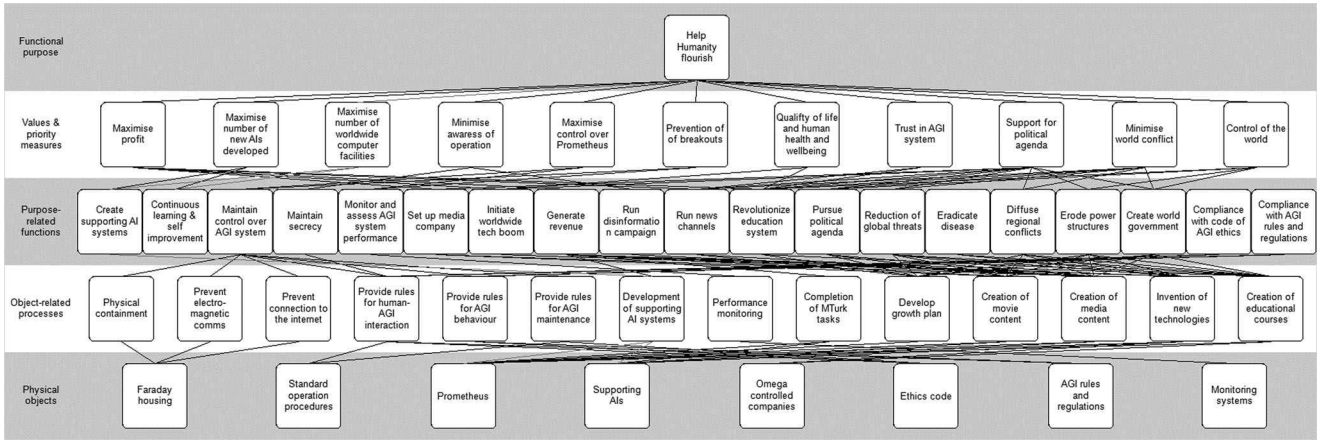

FIGURE 1 Artificial general intelligence design abstraction hierarchy example with example control objects, processes, and functions

These analyses will enable the identification and subsequent design of requisite control mechanisms. Importantly, based on an initial WDA, the resulting control mechanisms would be compatible with the AGI system's stated functional purposes, values, and priority measures, and purpose-related functions.

4.5 | SOCA

The SOCA phase can be used to identify how functions and associated strategies are distributed amongst agents (both human and technological) within the system. It also identifies how these agents can potentially communicate and cooperate (Vicente, 1999). The aim is to determine an optimum function allocation profile for the system in question. The SOCA process uses the outputs from the earlier CWA phases to identify what human and nonhuman agents currently do, and how functions, decisions, and strategies could potentially be allocated differentially, and in a dynamic manner.

SOCA has many potential applications in AGI design. For example, with AGI it seems important to establish not only who could do what, but also who should do what in light of moral, ethical and human needs (Hancock, 2009). Whilst a superintelligent AGI such as Prometheus might seemingly be able to do everything eventually, a degree of careful thought is required to determine what the optimum allocation of functions between humans and AGI is, to maintain human health, wellbeing, and safety. For example, organizations could use SOCA to establish which jobs they should allocate to AGI, and which should remain within the human domain. In the case of Prometheus, the AGI was responsible for developing a detailed "step-by-step growth plan" (Tegmark, 2017, p. 12). Ostensibly, a SOCA would likely have recommended that such planning should either be driven by human operators, or at least be undertaken in collaboration between the AGI and its human colleagues. Whilst such an analysis may invariably show that AGI can easily replace most human functions, the SOCA phase will be useful for identifying the optimum allocation of functions which minimizes risk and best facilitates human health and wellbeing.

SOCA would be useful to support the design of the standard operating procedures required to dictate how an AGI system such as Prometheus would work. Enacting this process would enable the Omega Team to identify what roles they need to continue to fulfill in order for the AGI to work efficiently but also safely and within control. In Tegmark's scenarios, Prometheus is given too much latitude for behavior and it undertakes too many functions without sufficient human intervention or authority. A SOCA could prevent this, identifying key roles that the Omega Team ought to have kept to maintain control over the Prometheus system.

4.6 | Designing AGI controls with STAMP

It goes without saying that all AGI systems will operate within a wider sociotechnical system. Initially Prometheus is implemented within a broader AI development company under the control of the Omega team. Presumably the company itself also operates within a broader system of regulations, design standards, and government control. A key contribution that systems HFE methods can make here is to help understand the broader sociotechnical system in which different AGIs are to be employed and to support the design of the control and feedback mechanisms required to maintain safety and minimize risk during AGI system design and operation. In the case of Prometheus, this form of analysis would go beyond the Omega Team and parent company to consider the roles and responsibilities of others in controlling AGI, including government, regulators, and internet infrastructure designers and managers to but a few.

As discussed earlier, it is widely acknowledged that a failure to implement appropriate controls could have catastrophic consequences (Amodei et al., 2016; Bostrom, 2014; 2017; Omohundro, 2014; Steinhardt, 2015). However, during Prometheus's development phase, AGI systems did not yet exist, so it would be extremely difficult to identify what controls will be required to manage its behavior. This represents an envisioned world problem and is one that HFE professionals have also previously faced. Thus, another systems HFE approach that can help here is the System Theoretic Accident Model and Process (STAMP; Leveson, 2004). According to STAMP, adverse events occur when interactions between systems components are not controlled through managerial, organizational, physical, operational, and manufacturing-based controls. The risks associated with AGI should therefore be managed through a hierarchy of controls and feedback mechanisms, and adverse events associated with AGI will result when the behavior of AGI systems is not adequately controlled. STAMP therefore views AGI safety as an issue of control and one that should be managed through a control structure that has the goal of enforcing constraints on AGI systems, their designers, and their managers.

The first phase of STAMP analyses involves building a control structure to describe the control relationships that exist between agents and organizations during both system design and operation. A generic control structure is presented in Figure 2. The control structure incorporates a series of hierarchical system levels and describes the agents and organizations that reside at each of these levels. Control (downward arrows) and feedback mechanisms (upward arrows) are included to show what controls are enacted down the hierarchy and what information about the status of the system is sent back up the hierarchy. A first important distinction to make with regard to AGI systems are that there are risks that require management both during the design and operation of AGI systems. Notably, much discussion in the literature relates to the requirement for controls when designing AGI systems (e.g., Everitt et al., 2018); however, there is comparatively little focus on the controls required during actual AGI system operation.

We developed an example control structure for AGI system operation based on the controls identified in the abstraction hierarchy (Figure 1) and our previous experiences in developing detailed

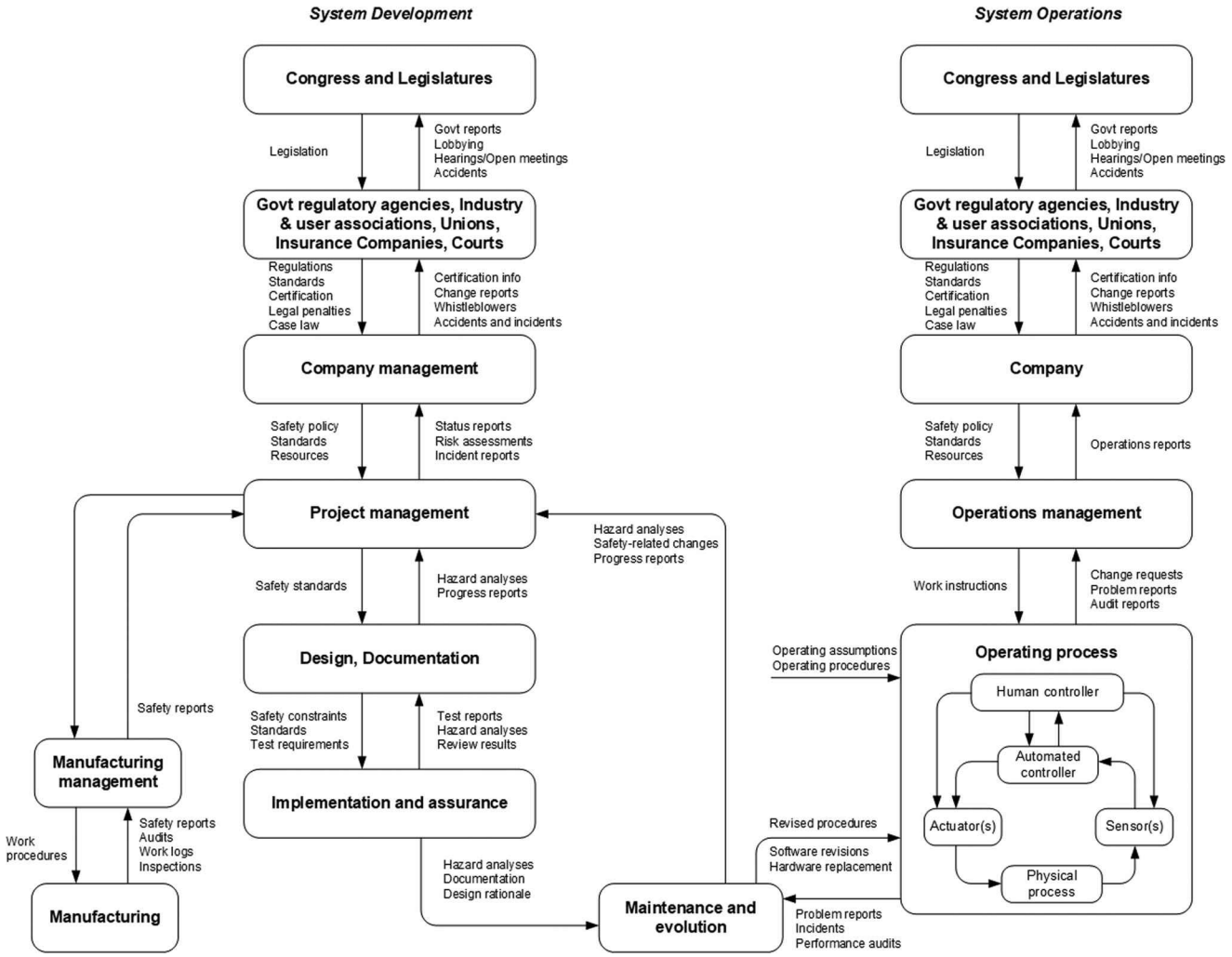

FIGURE 2 Generic control structure. *Source*: Adapted from Leveson (2004)

control structure models (e.g., Salmon & Read, 2019). The example AGI control structure is presented in Figure 3.

As shown in Figure 3, various forms of control will likely be required to ensure that AGI systems operate safely. In addition, various stakeholders will share the responsibility for AGI safety. For example, regulation at level 2 would be used to impose controls around how AGI can be used. Similarly, in a work context, at Level 3 organizations would presumably have standard operating procedures which dictate how the AGI should work and how human workers should interact with it. Finally, at Level 1 Federal and State governments would impose control on the level below (government agencies, industry associations, user groups, and the courts) through legislation. In Tegmark's scenario, it is notable that Prometheus was developed in secret, and disinformation campaigns are used to maintain secrecy. As such, many of the controls described above were circumvented.

The dashed arrows pointing upwards represent feedback mechanisms whereby agents and organizations provide information regarding the status of the system to the levels above. For example, "Government reports" are a feedback mechanism provided by Level 2 (government agencies, industry associations, user groups, and the courts) to Level 1 (parliament and legislatures). Feedback mechanisms exist between adjacent levels of the control structure (shown by straight dashed arrows) and also between nonadjacent levels (shown by curved dashed arrows). Of course, we acknowledge that in practice, such feedback mechanisms do not always function optimally (Hancock, 2019). This is a short-fall that is vital for HFE to address for successful implementation of AGI.

Arguably then, there are three sets of controls that require development and testing immediately:

1. the controls required to ensure AGI system designers and developers create safe AGI systems;
2. the controls that need to be in-built into the AGIs themselves, such as "common sense," morals, operating procedures, decision-rules, etc; and
3. the controls that need to be added to the broader systems in which AGI will operate, such as regulation, codes of practice, standard operating procedures, monitoring systems, and infrastructure.

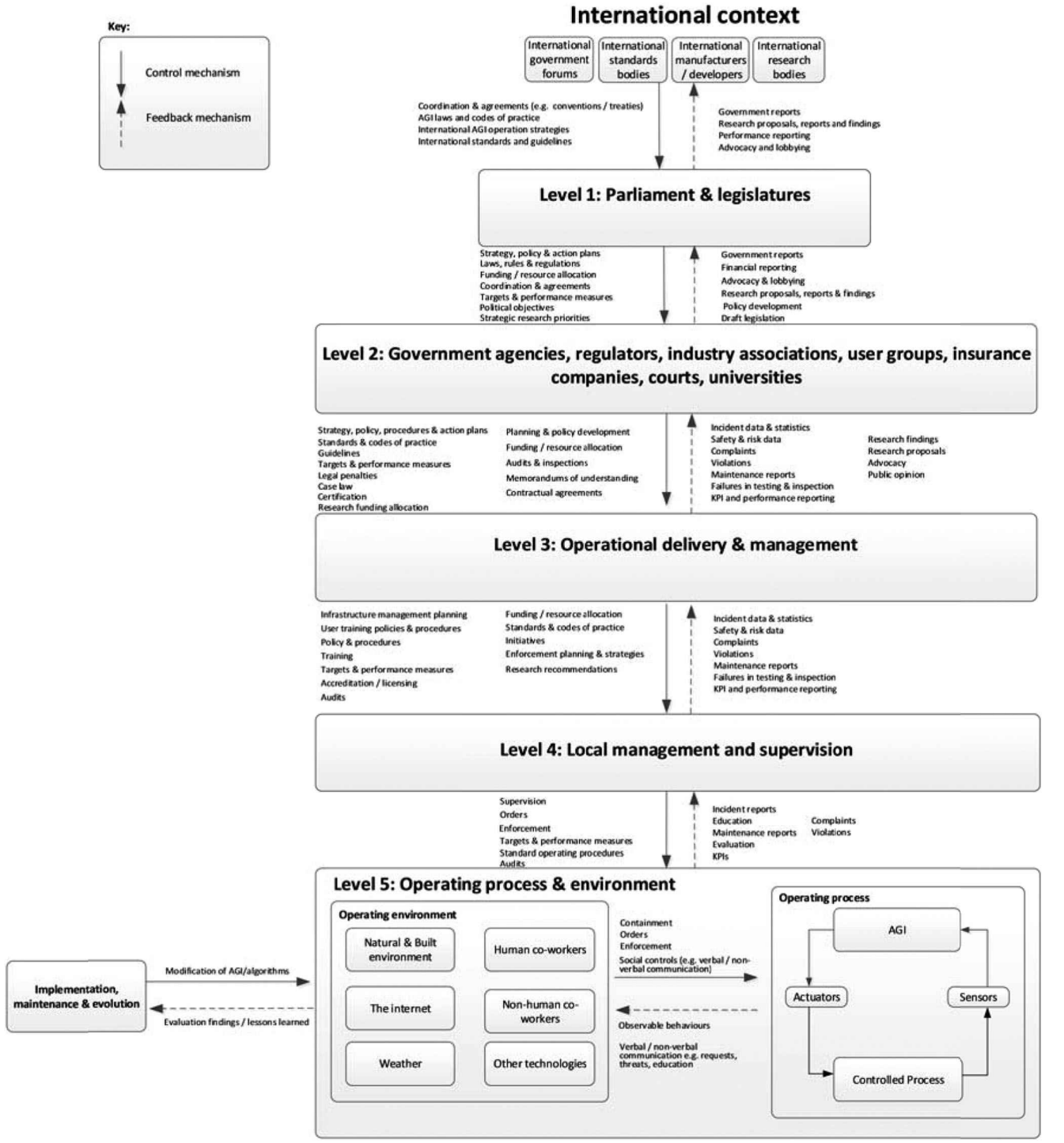

FIGURE 3 Example artificial general intelligence system operation control structure

The first two sets of controls are currently the primary focus of the AI community; however, the third set is where HFE can add the most value. With regard to Prometheus, the control structure component of STAMP could have been used to design the controls required to ensure that diffusion scenarios could have been prevented. Achieving this involves developing models of existing AI system control structures, and then identifying where new controls were required for AGI systems. Importantly, stakeholders from all levels of the system would need to be involved in the analysis and design activities to ensure that appropriate controls were built across the system, as opposed to only at the company level. For example, such an approach would enable governments and regulators to develop the controls required to prevent the covert development of AGI systems (as was the case with Prometheus).

TABLE 4 Example AGI STPA analysis

Control mechanism	Control failure	Consequences	Design requirements
Levels 1–2: Regulation	Inappropriate, ineffective, or missing control hazards: Appropriate AGI regulatory systems are not in place	Tech companies are able to develop and implement AGI systems without appropriate regulation	Pro-active development and formal evaluation of AGI regulatory system before development of first operational AGI system
	Control action provided too late: Regulation is developed and implemented after AGI systems are already developed and released	Tech companies are able to develop and implement AGI systems without appropriate regulation	Pro-active development and formal evaluation of AGI regulatory system before development of first operational AGI system
Levels 3–4: Supervision	Inappropriate, ineffective, or missing control hazards: Training provided to human supervisors does not cover the range of potential breakout scenarios/avenues	AGI is able to remove itself from control systems as human supervisor does not understand all of the potential ways in which the AGI can breakout Training is initially effective however	Training for AGI system controllers/supervisors/co-workers should incorporate various breakout scenarios and appropriate responses
	Control action stopped too soon: Training is provided once but not repeated annually and so does not keep pace with advances in capability of AGI	AGI systems soon self-improve and identify new ways of breaking out not considered in training	Trainees should understand the risks associated with AGI system breakouts as well as the range of different ways in which AGI systems can break out
			Regular review, update and delivery of training programs based on advances in AGI system capabilities
Levels 4–5: Supervision	Inappropriate, ineffective, or missing control hazards: AGI persuades human supervisor to connect it to a personal computer	AGI is able to connect to internet and break out from existing control systems	System should include hard engineered controls as well as human social controls

Abbreviations: AGI, artificial general intelligence; STPA, Systems Theoretic Process Analysis.

4.7 | Proactive AGI risk assessment

The Systems Theoretic Process Analysis (STPA; Leveson, 2017) method is a risk assessment method that is also used as part of STAMP to forecast instances where control and feedback mechanisms could potentially fail and develop appropriate risk controls. STPA works by considering each of the control and feedback mechanisms described in the control structure along with a control/feedback failure taxonomy comprising the following four failure modes (Leveson, 2011):

1. control or feedback action is not provided or followed;
2. an unsafe control or feedback action is provided;
3. control or feedback action is provided too early or too late;
4. control or feedback action is stopped too soon or applied too long.

The output therefore includes a description of potential control and feedback failures and their consequences which is used to support the identification, development, and implementation of appropriate risk controls.

Once the relevant stakeholders had devised a prototype control structure that they were satisfied with, STPA could therefore be used to forecast different instances in which such controls might fail or be circumvented (see Table 4 for examples). This enables stakeholders to strengthen control and feedback mechanisms and design and add new ones where necessary. Importantly, this control analysis and design work could occur years before Prometheus (or any form of AGI) was released. This would ensure that the appropriate controls had already been developed, tested, and implemented.

5 | CONCLUSION

The aim of this article was to discuss the potential role of different HFE methods in the design, implementation, and operation AGI systems. Our consideration of HFE methods leads to the conclusion that methods from fifteen distinct categories can, and should be, used to support AGI system design, implementation, and operation. Notably, many of the concerns regarding AGI systems relate to the capacity of humans, organizations, and broader sociotechnical systems to establish and sustain control over them. It is apparent that HFE methods are suited to identifying, designing, and testing not only AGI system controls, but also AGI systems themselves. In this respect, systems HFE methods, such as CWA and STAMP have important and demonstrable potential to help minimize the risks associated AGI.

To close, we can ask what is required to facilitate the contribution of HFE in AGI system design? First and foremost, despite the fact that AGI systems are not yet operational, applications of the methods discussed are required immediately. For example, STAMP could be used to analyze existing control structures used in different domains to manage new technologies to identify where they are or are not suitable for AGI systems and where new forms of control are required. Such knowledge would enable work on the development of AGI controls to begin now, and not after the fact as we have seen in the case of autonomous vehicles and other burgeoning technical innovations (Hancock, 2019; Salmon, 2019).

Second, HFE researchers and practitioners need to be involved in multi-disciplinary AGI design and development teams and programs. Currently, the presence of HFE researchers and practitioners within such multi-disciplinary programs appears to be sparse at best. Indeed, it has been noted that HFE practitioners often find it difficult to work with other professions, particularly those who are focussed on problems not typically considered to be HFE oriented (Thatcher et al., 2018). It is critical then that HFE connects with those other parties involved in the design of AGI systems. As AGI concepts and prototypes mature, it will become ever harder to embed HFE into AGI life cycles. We have repeatedly seen examples of this with technologies in areas such as defence (Stanton et al., 2010) and transport (Hancock, 2017; Salmon, 2019). We in HFE have to learn the lessons required to prevent a similar state-of-affairs with AGI.

Third, as shown in Table 2 there is an evident and immediate line of progress in refining some of our HFE theories and methods to ensure that we are better equipped to cope with AGI system design, analysis, and modeling. Questions can be legitimately raised regarding the applicability of our human-centered models to nonhuman superintelligent AGI systems. For example, concepts such as situation awareness, decision making, and cognitive workload will require redefinition for the AGI context. Likewise, methods such as cognitive task analysis, situation awareness assessment, mental workload assessment, and usability assessment will require refinement to enable their use with nonhuman participants.

Fourth and finally, other issues that have long troubled our discipline require resolution as HFE looks to influence AGI systems (Salmon, 2019). HFE practitioners need to more clearly demonstrate how the outputs of HFE analyses can inform design (Norros, 2014; Read et al., 2018) and specifically AGI system design. Whilst computational modeling is becoming more popular in HFE (Read et al., 2020), better methods of prediction are required (Moray, 2008), both for human and system performance (Salmon & Read, 2019). The gap between research and practice (Salmon et al., 2020; Shorrock & Williams, 2016) has to be closed, especially since the methods we have discussed in this article are more commonly used by researchers than they are by practitioners (Salmon et al., 2017). Perhaps most importantly, HFE has to be far better embedded throughout the design life cycles of new technologies and systems everywhere, not just in AI and AGI (Stanton et al., 2010). There is no doubt that our contribution to AGI systems can be facilitated by working through some of the issues that have long troubled our discipline (Salmon, 2019).

With AI, as Hancock (2019) has asserted, the horse has bolted, and HFE is again trying to catch up. With AGI, that horse has not yet bolted, but it is chomping at the bit to do so. Therefore, the next decade represents a critical period for HFE in this context. There is an opportunity to use HFE to help create safe and efficient AGI systems that can have far reaching benefits to society and all of humanity. At the same time, a business as usual approach could lead to the extinction of the human race. We believe that HFE researchers and practitioners are listening, as well as those that are leading the development of AGI systems. Now is the time to bring the two communities together.

ORCID


Paul M. Salmon http://orcid.org/0000-0001-7403-0286

REFERENCES

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. *AI*, 1–29.

Annett, J., Duncan, K. D., Stammers, R. B., & Gray, M. J. (1971). *Task analysis* (Department of Employment Training Information Paper 6). HMSO, London.

Bisantz, A. M., & Burns, C. M. (2008). *Applications of cognitive work analysis*. CRC Press.

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

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

Bostrom, N. (2017). Strategic implications of openness in AI development. *Global Policy*, *8*(2), 135–148.

Brooke, J. (1996). SUS—A quick and dirty usability scale. In P. Jordan, B. Thomas, I. McLelland, & B. A. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194).

Brundage, M., Avin, S., Clark, J., et al., (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. http://www.maliciousaireport.com

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*, 9–10.

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.

Embrey, D. E. (1986) SHERPA: A systematic human error reduction and prediction approach. Proceedings of the International Topical Meeting on Advances in Human Factors in Nuclear Power Systems, Knoxville, Tennessee American Nuclear Society, La Grange Park, IL.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65–84.

Everitt, T., Lea, G., & Hutter, M. (2018). AGI safety literature review. *IJCAI*. arXiv.

Future of Life Institute (FLI). (2018). Benefits and risks of artificial intelligence. https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/. Accessed January, 2019.

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling AI: Human perils. *Computer Law and Security Review*, *32*(5), 749–758.

Hancock, P. A. (2009). *Mind, machine, and morality*. Ashgate.

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

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

Hancock, P. A., & Matthews, G. (2018). Workload and performance: Associations, Insensitivities, And Dissociations. *Human Factors*, *61*(3), 374–392.

Hancock, P. A., Nourbaksh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences of the United States of America*, *166*(16), 7684–7691.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research". In P. A. Hancock, & N. Meshkati (Eds.), *Human mental workload* (pp. 5–39). Elsevier.

International Ergonomics Association. (2020). *Human factors/ergonomics*. https://iea.cc/what-is-ergonomics/

Jenkins, D. P., Stanton, N. A., Salmon, P. M., & Walker, G. H. (2009). *Cognitive work analysis: coping with complexity*. Ashgate.

Kaplan, A., & Haenlein, M. (2018). 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.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 30, pp. 576–580). SAGE Publications.

Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Gerald Duckworth & Co. Ltd.

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

Leveson, N. G. (2011). Applying systems thinking to analyze and learn from events. *Safety Science*, *49*, 55–64.

Leveson, N. G. (2017). Rasmussen's legacy: A paradigm change in engineering for safety. *Applied Ergonomics*, *59*(Part B), 581–591.

Moray, N. (2008). The good, the bad, and the future: On the archaeology of ergonomics. *Human Factors*, *50*(3), 411–417.

National Highway Traffic Safety Administration. (2016). *Office of defects investigation: PE 16-007*. https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

National Transportation Safety Board. (2018). *Preliminary report highway HWY18MH010*. https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

Navarro, J. (2019). A state of science on highly automated driving. *Theoretical Issues in Ergonomics Science*, *20*, 336–396.

Norros, L. (2014). Developing human factors/ergonomics as a design discipline. *Applied Ergonomics*, *45*(1), 61–71.

Omohundro, S. (2014). Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 303–315.

Pennachin, C., & Goertzel, B. (2007). Contemporary approaches to artificial general intelligence. In B. Goertzel, & C. Pennachin (Eds.), *Artificial general intelligence* (pp. 1–30). Springer.

Rasmussen, J. (1986). *Information processing and human-machine-interaction—An approach to cognitive engineering*. North-Holland.

Ravden, S., & Johnson, G. (1989). *Evaluating usability of human-computer interfaces: A practical method*. Conference on Human Factors in Computing Systems. Ellis Horwood Ltd.

Raven, P. G. (2017). Telling tomorrows: Science fiction as an energy futures research tool. *Energy Research & Social Science*, *31*, 164–169.

Read, G. J. M., Salmon, P. M., Lenne, M. G., & Goode, N. A. (2018). A sociotechnical design toolkit for bridging the gap between systems-based analyses and system design. *Human Factors and Ergonomics in Manufacturing and Service Industries*, *28*(6), 327–341.

Read, G. J. M., Salmon, P. M., Thompson, R., & McClure, R. (2020). Simulating the behaviour of complex systems: Computational modelling in ergonomics. *Ergonomics*, *63*, 931–937.

Salmon, P. M. (2019). The horse has bolted! Why human factors and ergonomics has to catch up with autonomous vehicles (and other advanced forms of automation). Invited Commentary on Hancock (2018) "Some Pitfalls in the Promises of Automated and Autonomous Vehicles". *Ergonomics*.

Salmon, P. M., Carden, T., Hancock, P. (In Press). Putting the humanity into inhuman systems: how human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *Human Factor and Ergonomics in the Manufacturing and Service Industries*, Accepted for publication 20th November 2020.

Salmon, P. M., Hancock, P., & Carden, T. (2019). To protect us from the risks of advanced artificial intelligence, we need to act now. *The Conversation*, 502–504.

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., Walker, G. H., Read, G. J. M., Goode, N., & Stanton, N. A. (2017). Fitting methods to paradigms: are ergonomics methods fit for systems thinking? *Ergonomics*, *60*(2), 194–205.

Sawyer, B. D., & Hancock, P. A. (2018). Hacking the human: The prevalence paradox in cybersecurity. *Human Factors*, *60*(5), 597–609.

Shorrock, S. T., & Williams, C. (2016). Human factors and ergonomics methods in practice: Three fundamental constraints. *Theoretical Issues in Ergonomics Science*, *17*(5), 1–15.

Stanton, N. A., Jenkins, D. P., Salmon, P. M., Walker, G. H., Rafferty, L., & Revell, K. (2010). *Digitising command and control: A human factors and ergonomics analysis of mission planning and battlespace management*. Ashgate.

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

Stanton, N. A., Salmon, P. M., Walker, G., & Stanton, M. (2019). Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. *Safety Science*, *120*, 117–128.

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

Steinhardt, J. (2015). *Long-term and short-term challenges to ensuring the safety of AI systems*. https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Vintage Books.

Thatcher, A., Waterson, P., Todd, A., & Moray, N. (2018). State of science: Ergonomics and global issues. *Ergonomics*, *61*(2), 197–213.

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

Walker, G. H., Salmon, P. M., Bedinger, M., & Stanton, N. A. (2017). Quantum ergonomics: Shifting the paradigm of the systems agenda. *Ergonomics*, *60*(2), 157–166.

Yudkowsky, E. (2012). Friendly artificial intelligence. In A. H. Eden, J. H. Moor, J. H. Soraker, & E. Steinhart (Eds.), *Singularity hypotheses: A scientific and philosophical assessment* (pp. 181–195). Springer.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In N. Bostrom, & M. M. Ćirković (Eds.), *Global catastrophic risks* (pp. 308–345). Oxford University Press.

How to cite this article: Salmon PM, Carden T, Hancock PA. Putting the humanity into inhuman systems: How human factors and ergonomics can be used to manage the risks associated with artificial general intelligence. *Hum Factors Man.* 2020;1–14. https://doi.org/10.1002/hfm.20883