

## The risks associated with artificial general intelligence

McLean, Scott; Read, Gemma J.M.; Thompson, Jason; Baber, Chris; Stanton, Neville A.; Salmon, Paul M.

DOI:

[10.1080/0952813X.2021.1964003](https://doi.org/10.1080/0952813X.2021.1964003)

License:

Creative Commons: Attribution-NonCommercial-NoDerivs (CC BY-NC-ND)

*Document Version*

Publisher's PDF, also known as Version of record

*Citation for published version (Harvard):*

McLean, S, Read, GJM, Thompson, J, Baber, C, Stanton, NA & Salmon, PM 2021, 'The risks associated with artificial general intelligence: a systematic review', *Journal of Experimental and Theoretical Artificial Intelligence*. <https://doi.org/10.1080/0952813X.2021.1964003>

[Link to publication on Research at Birmingham portal](#)

### General rights

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

### Take down policy

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.

If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.



## The risks associated with Artificial General Intelligence: A systematic review

Scott McLean, Gemma J. M. Read, Jason Thompson, Chris Baber, Neville A. Stanton & Paul M. Salmon

To cite this article: Scott McLean, Gemma J. M. Read, Jason Thompson, Chris Baber, Neville A. Stanton & Paul M. Salmon (2021): The risks associated with Artificial General Intelligence: A systematic review, Journal of Experimental & Theoretical Artificial Intelligence, DOI: [10.1080/0952813X.2021.1964003](https://doi.org/10.1080/0952813X.2021.1964003)

To link to this article: <https://doi.org/10.1080/0952813X.2021.1964003>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 13 Aug 2021.



Submit your article to this journal [↗](#)



Article views: 2906



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

# The risks associated with Artificial General Intelligence: A systematic review

Scott McLean <sup>a</sup>, Gemma J. M. Read<sup>a</sup>, Jason Thompson<sup>a,b</sup>, Chris Baber<sup>c</sup>, Neville A. Stanton<sup>a</sup> and Paul M. Salmon <sup>a</sup>

<sup>a</sup>Centre For Human Factors And Sociotechnical Systems, University Of The Sunshine Coast, Sippy Downs, Australia; <sup>b</sup>Transport, Health and Urban Design (Thud) Research Lab, Melbourne School of Design, The University of Melbourne, Parkville, Victoria, Australia; <sup>c</sup>School Of Computer Science, University Of Birmingham, Birmingham, UK

## ABSTRACT

Artificial General intelligence (AGI) offers enormous benefits for humanity, yet it also poses great risk. The aim of this systematic review was to summarise the peer reviewed literature on the risks associated with AGI. The review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Sixteen articles were deemed eligible for inclusion. Article types included in the review were classified as philosophical discussions, applications of modelling techniques, and assessment of current frameworks and processes in relation to AGI. The review identified a range of risks associated with AGI, including AGI removing itself from the control of human owners/managers, being given or developing unsafe goals, development of unsafe AGI, AGIs with poor ethics, morals and values; inadequate management of AGI, and existential risks. Several limitations of the AGI literature base were also identified, including a limited number of peer reviewed articles and modelling techniques focused on AGI risk, a lack of specific risk research in which domains that AGI may be implemented, a lack of specific definitions of the AGI functionality, and a lack of standardised AGI terminology. Recommendations to address the identified issues with AGI risk research are required to guide AGI design, implementation, and management.

## ARTICLE HISTORY

Received 20 January 2021  
Accepted 28 July 2021

## KEYWORDS

Artificial General Intelligence; artificial intelligence; risk; existential threat; safety

## Introduction

Artificial General Intelligence (AGI) is the next generation of artificial intelligence (AI), which is expected to exceed human intelligence in every aspect (Barrett & Baum, 2017; Bostrom, 2014; Torres, 2019). AGI will extend upon AI, or Artificial Narrow Intelligence (ANI) systems, which are in widespread use today. For example, current ANI systems include Google's DeepMind, Facebook's facial recognition technology, Apple's 'Siri', Amazon's Alexa, and Tesla's and Uber's self-driving vehicles (Kaplan & Haenlein, 2019; Naudé & Dimitri, 2020; Stanton et al., 2020). ANI systems use deep learning algorithms to analyse large volumes of data to make predictions regarding behaviour in specific tasks (LeCun et al., 2015; Naudé & Dimitri, 2020). As such, an ANI's intelligence is task specific (or narrow) and cannot transfer to other domains with unknown and uncertain environments in which they have not been trained (Firt, 2020).

In contrast, an AGI would possess a different level of intelligence (Bostrom, 2014), which has previously been defined as an agent's ability to achieve goals in a wide range of environments (Legg

& Hutter, 2006), and the ability to achieve complex goals in complex environments (Goertzel, 2006). Whilst current ANI systems have typically been used as tools to support human behaviours, an AGI system would be an autonomous agent that can learn in an unsupervised manner (Firt, 2020; Torres, 2019). Whilst, AGI does not currently exist, it is expected to arrive sometime this century (Müller & Bostrom, 2016).

Although ANI systems such as Uber's automated vehicles can create safety risks (Stanton et al., 2019), they do not, at present, pose a significant threat to humanity (Bentley, 2018). This is not the case with AGI, with many scholars discussing potential existential threats (Salmon et al., 2021). The risks associated with AGI are generated by the challenge of controlling an agent that is substantially more intelligent than us (Baum, 2017). The exponential rate at which technology is advancing, such as in the areas of computing power, data science, neuroscience, and bioengineering, has led many scholars to believe that an intelligence explosion will be reached in the near future (Kurzweil, 2005; Naudé & Dimitri, 2020). An intelligence explosion would see AI exceed human-level intelligence (Chalmers, 2009). At this point, which is estimated to occur between 2040 to 2070 (Baum et al., 2011; Müller & Bostrom, 2016), it is hypothesised that an AGI will have the capability to recursively self-improve by creating more intelligent versions of itself, as well as altering their pre-programmed goals (Tegmark, 2017). The emergence of AGI could bring about numerous societal challenges, from AGI's replacing the workforce, manipulation of political and military systems, through to the extinction of humans (Bostrom, 2002, 2014; Salmon et al., 2021; Sotala & Yampolskiy, 2015). Given the many known and unknown risks regarding AGI, the scientific community holds concerns regarding the threats that an AGI may have on humanity (Bradley, 2020; Yampolskiy, 2012). These concerns include malevolent groups creating AGI for malicious use, as well as catastrophic unintended consequences brought about by apparently well-meaning AGI's (Salmon et al., 2021). There is much scepticism among experts as to whether AGI will ever eventuate, and responses to the AGI debate are broad and range from doing nothing, as an AGI may never be created (Bringsjord et al., 2012), to the extremes of allowing AGI to destroy humanity and take our place in an evolutionary process (Garis, 2005).

Despite the scepticism, Baum (2017) identified 45 active AGI research and development projects, including Deepmind, Open AI, GoodAI, CommAI, SingularityNET (Baum, 2017; Torres, 2019). If AGI is successfully developed, it is argued that there will be only one chance to ensure that the design, implementation and operation of AGI is appropriately managed, as rapid advances will immediately render the initial AGI obsolete (Bostrom, 2014). This is highly problematic when considering risk management, as the initial risk controls may also be ineffective as the AGI redesigns and self-improves. As such, there is an urgent need to understand and develop appropriate risk controls now, to ensure the creation of safe AGI's and continued and effective management of associated risks as they develop (Salmon et al., 2021; Sotala & Yampolskiy, 2015). Salmon et al. (2021), for example, recently outlined a research agenda designed to ensure that appropriate design and risk management methods are immediately embedded in AGI design. Despite calls such as this, the extent to which the research community is actively exploring the risks associated with AGI in scientific research is not clear (Baum, 2017). Moreover, the specific nature of the risks associated with AGI is not often made clear, with discussions focusing more on general existential threats such as AGI systems deciding that humans are no longer required (Bostrom, 2014). The literature on the risks associated with AGI has grown substantially within the last decade, and includes numerous books, government and academic white papers, website blogs and articles, and conference proceedings, among others (Sotala & Yampolskiy, 2015). However, the extent to which this has translated into formal scientific studies exploring the risks associated with AGI is not clear. The purpose of this systematic review is therefore to examine and report on the peer reviewed scientific literature that has specifically investigated the risks associated with AGI. The intention was to determine the level of scientific inquiry in this area and to identify specifically what forms of risk are being explored. As such the specific research question investigated for the current systematic review was: *What are the risks associated with Artificial General Intelligence?*

## Methods

### Protocol

The current systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009). PRISMA includes an evidence-based set of reporting guidelines to ensure that the literature review process is comprehensive, replicable and transparent. Data were extracted from each included study to understand the current state of the literature relating to the Research Question. For the current systematic review, we have assumed that AGI will be created. For clarity, in the current review, the term AGI is used to represent non-human agents with human level and above intelligence.

### Electronic search

The literature search was completed on the 26<sup>th</sup> of August 2020. Literature was searched across three databases: SCOPUS, Web of Science (WoS), and Association for Computational Machinery (ACM). The citation software Endnote X9 was used to facilitate the searching and screening process. Based on the research question, Boolean search terms were developed, and appropriate limiters applied. The Boolean search terms and filters were applied to ensure that only relevant studies were included in the initial search. The search terms and filters applied to all three databases was: 'Artificial general intelligence' OR 'artificial superintelligence' OR 'strong AI' OR 'technological singularity' OR 'technological explosion' AND 'safety' OR 'risk\*' OR 'Danger\*' OR 'Threat\*' OR 'accident\*' OR 'security' OR 'unintended behaviour' OR 'societal impacts' Or 'Value\*' (English: Article). All retrieved articles' title and abstract were assessed based on the eligibility criteria. Articles that did not align with the eligibility criteria were excluded from the review.

### Eligibility criteria

A set of inclusion and exclusion criteria were developed to support the identification of studies relevant to the research question.

### Inclusion criteria

Articles were included in the review if they were:

- Focused on AI systems with human level intelligence and above (e.g., artificial general intelligence and artificial superintelligence);
- Focused on risks associated with AGI;
- Were published in peer reviewed journals; and
- Were published in English.

### Exclusion criteria

Articles were excluded from the review where they:

- Focused only on artificial intelligence (AI) or artificial narrow intelligence (ANI);
- Focused only on AGI algorithms/engineering/architecture/programming;
- Focused only on the predicted date of arrival of human level intelligence and above;
- Conference proceedings, grey literature, reviews, arXiv articles, journal editorials, books; and
- Were commentaries on the different viewpoints related to human level intelligence and above.

## **Data extracted**

The following information was extracted from each paper:

- Author(s) and year of publication;
- Domain/topic (e.g., defence, transport, healthcare, general (non-specific));
- AGI specification/functionality;
- Analysis method;
- Risks identified; and
- Risk controls/risk management strategies.

## **Results**

In total, 136 articles from the three databases were identified; Scopus (n = 61), WoS (n = 58), and ACM (n = 17). Following the removal of duplicates (n = 39), the title and abstract of 97 articles were assessed which resulted in 80 articles excluded for not meeting the eligibility criteria. This resulted in 17 full text articles selected for eligibility assessment, of which six did not meet eligibility criteria and were excluded. A further 18 potential articles were identified by screening the article titles within in the bibliographies of the eligible articles, and full texts were assessed for eligibility. This process resulted in the inclusion of five articles. Altogether, this process resulted in the inclusion of 16 articles in the systematic review (Figure 1). Table 1 provides a list of studies included in the review.

### **AI specification/functionality**

Within the included articles, three described the specific functionality of the AGI. These included 12 different human professions, smart homes, and autonomous vehicles (Chen & Lee, 2019), a malicious AGI (Bradley, 2020), and AGI in the manufacturing, communication, and technology sectors (Narain et al., 2019). The remaining articles referred to generic AGI systems and did not describe a specific AGI functionality.

### **Analysis methods**

Within the included articles, multiple analysis methods were used to identify and understand the risks associated with AGI (Table 2). The most frequent type of analysis method included in the current systematic review were philosophical discussions (n = 8), in which arguments on the risk of AGI were developed based on existing theories and concepts. The remaining articles utilised a broad range of analysis methods within the context of the risks posed by AGI. Two articles assessed the capabilities of current risk management and legal standards and processes in relation to AGI, five articles applied different types of modelling. Finally, one study used surveys to obtain people's perception of the risk associated AGI across three domains (human professions, automated vehicles, smart homes).

### **Risk categories**

The risks identified in the articles can be broadly categorised as follows: AGI removing itself from the control of human owners/managers, AGIs being given or developing unsafe goals, development of unsafe AGI, AGIs with poor ethics, morals and values, inadequate management of AGI, existential risks (Table 3).

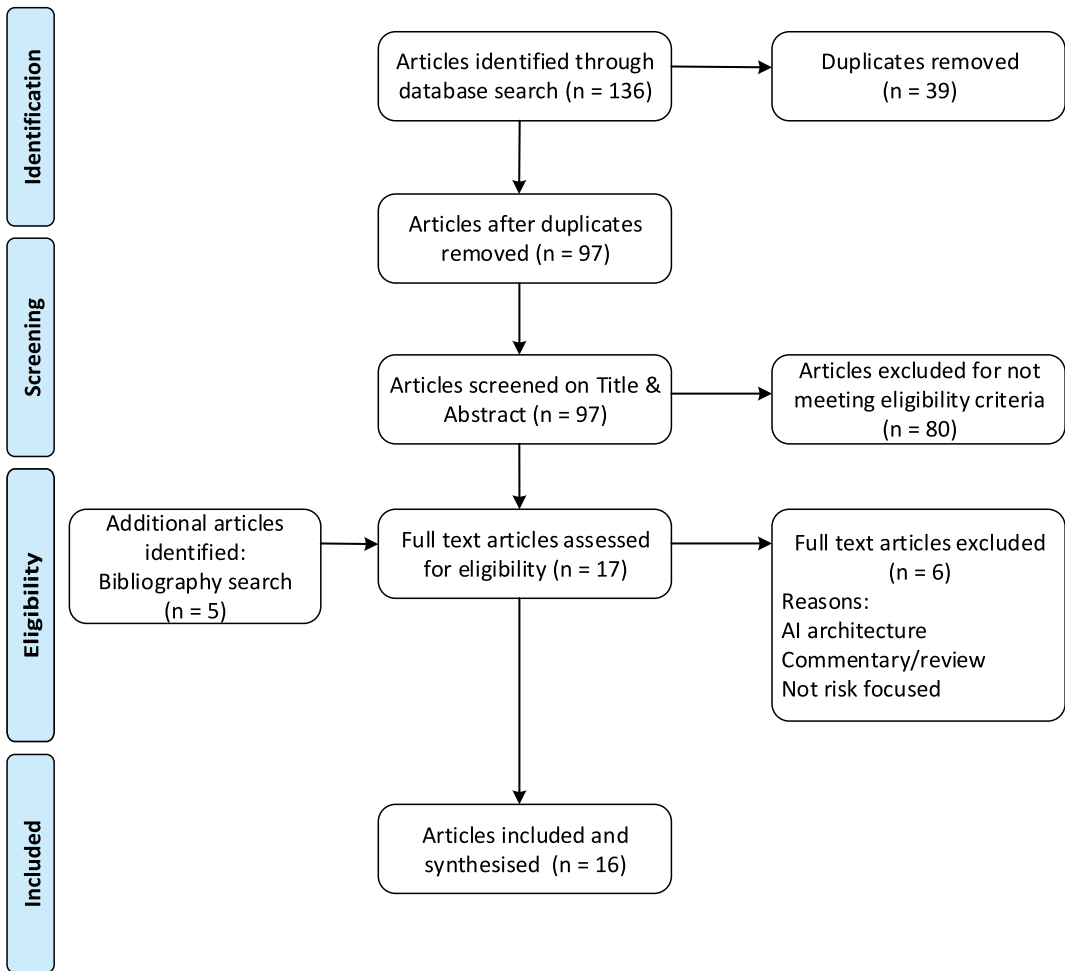


Figure 1. Systematic literature review process and results.

### Risk management and controls

Multiple forms of risk control were discussed across the included articles (Table 1). Specific controls involved the programming, development, and release of an AGI, increased data driven modelling, improved risk management processes, international regulations, government control, taxation of the AGI, and consultation with a wide range of experts, among others.

### Discussion

The purpose of this systematic review was to summarise the peer reviewed literature specifically investigating the risks associated with AGI. Whilst only a small set of articles were identified, several key findings can be taken from the review. A broad range of risks were identified, and encouragingly, a number of recommendations for requisite controls to manage the risks are discussed in the articles. However, apparent issues emerged regarding the current scope of peer reviewed literature on the risks posed by AGI. The following sections will discuss the key findings from the included articles, potential issues associated with current AGI research, and future research directions on the risks of AGI.

**Table 1.** Extracted data from the included articles.

Author(s) Publication year	Domain/topic	AGI Specification/functionality	Analysis method	Risk categories	Risk controls/risk management strategies
Armstrong et al. (2012)	General (nonspecific)	N/A	Philosophical discussion	AI confinement/control	An Oracle AI that is confined and can only answer questions rather being able to act.
Barrett and Baum (2017)	General (nonspecific)	N/A	Fault trees and influence diagrams	ASI take off Containment fails ASI goal safety	Modelling to help developers make better decisions to reduce the risk of catastrophic ASI; ASI research risk review boards; encourage research into ASI safety; enhance human capabilities; AI confinement; AI enforcement.
Baum et al. (2017)	General (nonspecific)	N/A	ASI risk estimate modelling on expert judgement	Containment fails AGI goal safety	Consider arguments and ideas from a wider range of experts; conduct formal expert surveys to elicit expert judgments of risk model parameters.; risk modelling towards identifying which opportunities are most effective at reducing the risk.
Boyles (2018)	Ethics, moral reasoning	N/A	Philosophical discussion	Survival of human values and norms Capability of machines for moral-reasoning, judgement, decision making	Creation of artificial moral agents; philosophical concepts such as moral reasoning, moral agency; issues related to AI research that may also be further studied by philosophers.
Bradley (2020)	Risk management	Malicious AI	Analysis of current risk management standards in relation to ASI. AI treachery threat model	Risk identification Assessment of likelihood Consequence bias	Recognition that current risk management is an incomplete approach to ASI; risk models to shift from static, anthropomorphic models to focus on data-driven models that measure intent, manage intent and prevent the treacherous turn.
Brundage (2014)	Machine ethics	N/A	Philosophical discussion	Morals and Ethics	No hope of machine ethics building on an existing technical core; adoption of a broad lens to analyse the inherent difficulty intelligent action and the complex social context in which humans and AGIs will find themselves.
Chen and Lee (2019)	Perceptions of the impact of AGI on humanity	Factory operator; translator; retail salesperson; tutor; accountant; news production staff; money management specialist; researcher; artist; surgeon; truck driver; home service robot; autonomous vehicle; smart home.	Questionnaire/survey	Social impact of AGI on 12 professions, AVs, and smart homes	Better communication of public perceptions of the benefits and concerns of specific AI applications to AI experts.

*(Continued)*



Table 1. (Continued).

Author(s) and Year	Topic	Key Concepts / Focus	Philosophical discussion	Risks to humanity	Conclusions / Recommendations
Goertzel and Pitt (2014)	General (nonspecific)	N/A	Philosophical discussion	Risks to humanity	Engineer the capability to acquire integrated ethical knowledge; provide rich ethical interaction and instruction, respecting developmental stages; develop stable, hierarchical goal systems; ensure that the early stages of recursive self-improvement occur relatively slowly and with rich human involvement; tightly link AGI with the Global Brain; foster deep, consensus-building interactions between divergent viewpoints; create a mutually supportive community of AGIs; encourage measured co-advancement of AGI software and AGI ethics theory; develop advanced AGI sooner not later. Humanity should be more optimistic about its long-term survival if we have convincing evidence for believing that both unfriendly AGI and the great filter are real, than if there was evidence for thinking that only one of these would lead to existential risk.
Miller (2019)	General (nonspecific)	N/A	Bayesian modelling	Dangers of unfriendly AGI	A gradual diffusion process; Threats pre-empted by long-term social measures; modelling efforts made to understand the extant perspective of the development of the high-technology diffusion.
Narain et al. (2019)	Management of ASI	Manufacturing; communication; and technology	Diffusion modelling, predator-prey models and hostility models.	Management and control of ASI	Introducing an intermediate outcome (e.g. second prize rather than one dominant AGI); using public procurement of innovation; taxing an AGI; addressing patenting by AI
Naudé and Dimitri (2020)	General (nonspecific)	N/A	Theoretical all-pay contest model	The race to develop AGI	An International Treaty to Regulate AI Research and Development; International Oversight over AI
Nindler (2019)	Legal	N/A	Analysis of current legal and institutional framework of the UN	Legal capabilities for the management of existential risks posed by AGI	Use of Force as a final possibility f;
Pueyo (2018)	Environmental/ social	N/A	Philosophical discussion	Environmental and social implications of superintelligence emerging in an economy shaped by neoliberal policies	Degrowth as a viable alternative; systemic change altering the motivations of economic action; changing values in firms, governments, and social movements to ease the change in individual values and reduce the risk of having people engaged in the development of undesirable forms of AI.

(Continued)

**Table 1.** (Continued).

Sotata and Gloor (2017)	General (nonspecific)	N/A	Philosophical discussion	Suffering risks (risks worse than extinction)	Alignment of human and AI values; Multi layered goal functions e.g. Fallback goals; separate superintelligence designs from ones that would contribute to suffering risk; identify the enabling factors of suffering; maximise the probability of the best worst-case scenario.
Torres (2019)	General (nonspecific)	N/A	Philosophical discussion	The race to develop AGI AGI control Dangers of AI denialism The AGI confinement Problem	All parties involved in the creation of AGI recognise the enormity of getting AGI wrong; need to cooperate and not cut safety corners to reach the AGI finish line first. Countermeasures against escape: preventing social engineering attacks; against system resource attacks and future threats; against external causes of escape; against information in-leaking (add); how to safely communicate with a superintelligence.
Yampolskiy (2012)	General (nonspecific)	N/A	Philosophical discussion		

## ***Findings from the studies reviewed***

### ***Types of analyses***

A range of analysis methods were identified in the included articles. These included philosophical discussions, various modelling approaches, and assessment of current standards and procedures in relation to AGI risk. One half of the included articles were philosophical discussions on the potential risks associated with AGI systems. These discussions centred upon AGI confinement, AGI control, machine values and morals, risk management, risks to humanity, and the race to develop AGI. Whilst philosophical viewpoints on the risks of AGI have provided compelling and ostensibly sound arguments, there is considerable disagreement between experts. For example, AI thought leaders Nick Bostrom and Ben Goertzel disagree on containment fails, human attempts to make AGI goals safe, and AGI not making its own goals safe, among others (Baum et al., 2017). Whilst the debate itself is not in the scope of this review, it is important to note that such philosophical debates have facilitated critical analyses of views on the potential risks associated with the creation of AGI. Given that little is known about the actual risks of AGI (Baum, 2017), philosophical discussions and thought experiments have been a necessary starting point to provide direction for AGI risk research.

Various data driven and theoretical models have previously been used to forecast the behaviour of emerging products and technologies, and will be necessary for estimating the risks of AGI (Narain et al., 2019). The modelling approaches identified in the reviewed articles included fault trees and influence diagrams, Bayesian modelling, diffusion modelling, predator–prey models, hostility models, theoretical all-pay contest model, risk estimate modelling and AI treachery threat model. The outcomes of the modelling studies in the current review have provided information to better understand AGI containment fails (Barrett & Baum, 2017; Baum et al., 2017), technology diffusion enabling a safer AGI (Narain et al., 2019), and the appropriateness of current risk management procedures to deal with the risks of AGI (Bradley, 2020). Even though AGI does not yet exist, these modelling approaches have demonstrated the capability to derive information to better inform decision making the risks associated with AGI.

The assessment of current risk management processes and legal capabilities demonstrated, that currently, they are not fit for purpose when it comes to AGI (Bradley, 2020; Nindler, 2019). For example, the current global standard for risk management (ISO 31,000:2018), and its specific risk assessment sections (risk identification, assessment of likelihood, and consequence) have critical shortcomings (Bradley, 2020). It was indicated that when applied to AGI, the current framework is vulnerable to unanticipated risks, and as such, alternative approaches to AGI risk management are required (Bradley, 2020). In addition, the United Nations (UN) institutional and legal capabilities to manage existential threats posed by AI require strengthening (Nindler, 2019). For example, there is a need for an international regulation on AGI research and development, and an international enforcement agency for safe AGI. It is worth noting that risk management processes and regulations have been found to be inadequate and lagging behind the development of ANI. A clear implication of this review is that the work required to ensure that risk management processes and regulatory frameworks are fit for purpose for AGI should begin now, regardless of whether AGI exists or not. As discussed by many of the leading scholars in the area, leaving this until the first AGI systems arrive is unacceptable and poses risk, itself.

### ***What are the risks and risk controls associated with AGI***

Numerous risks associated with AGI have been stated within the AGI literature, as well as controls to avoid the associated risks (Sotala & Yampolskiy, 2015). These risks and controls broadly fit within three categories i.e. societal, and external and internal constraints on AGI behaviour (Sotala & Yampolskiy, 2015). Societal risks and controls refer to, how as a society, risks are dealt with, for example, regulating AGI research. External constraints refer to the restrictions imposed on the AGI, for example, confinement of the AGI. Last, internal constraints are concerned with the design of the

AGI, such as building in safe motivations of AGI (Sotala & Yampolskiy, 2015). The risk categories and controls identified in the current review, correspond with these three categories. For example, the race to develop AGI, and the appropriateness of current risk management processes and legal frameworks accord with societal risks and controls associated with AGI risks. AGI containment or confinement fails fall within the external constraints imposed on the AGI. The design of AGI goal safety, and AGI ethics, morals, and values fit within the internal constraints imposed on AGI's. Whilst considerations and actions within each of the three categories (societal, external, and internal) may have merit in avoiding catastrophic AGI risk, they also suffer numerous limitations (Sotala & Yampolskiy, 2015). As such, integration of them may be necessary for limiting risk and achieving the desired outcomes of AGI. At present this appears difficult, for example, the majority of modelling approaches in the current review investigated risks associated with one of the three areas to avoid catastrophic risk. Only one of the reviewed studies incorporated societal constraints (review boards), external constraints (e.g., containment fails), and internal constraints (e.g., goal safety) to demonstrate potential pathways and avoidance of AGI catastrophe (Barrett & Baum, 2017). In the survey of active AGI projects by Baum (2017), the majority of projects had no identifiable engagement with safety, which led the authors to conclude that risk is not a primary focus of AGI research and development. As such, AGI risk researchers could benefit from adopting risk analysis methods from other scientific disciplines. For example, in the field of safety science, it has been shown that adverse events share a common causal network of contributory factors and relationships underpinning their aetiology (Salmon, Hulme et al., 2020).

### *Limitations of the AGI literature and future directions*

In the reviewed articles, there was a limited number of studies that focused on the risks to specific domains, as well as descriptions of specific functionality of the AGIs. The majority of studies reviewed were non-domain specific and focused on the general risks to humanity. However, specific domains identified in the eligible articles included autonomous vehicles, human professions, and smart homes (Chen & Lee, 2019), manufacturing, communication, and energy (Narain et al., 2019), law (Nindler, 2019), environment and social aspects (Pueyo, 2018). Two notable domains that did not feature in the review were research investigating the risks associated with defence and autonomous weapons systems, and healthcare. Given the obvious catastrophic risks associated with losing control of autonomous weapons systems to an AGI, an understanding of the potential risks is critical. AGI research with connections to the military is being conducted, in a 2017 survey of active AGI research and development projects, nine out of the total 45 active AGI projects had military connections (Baum, 2017). However, no military research was identified by the current review search strategy, which indicates two logical perspectives 1) the research may be confidential and not published, or 2) the research does not focus specifically on risk (Baum, 2017). In healthcare, we have witnessed how rapidly ANI has changed medical practice, for example, disease diagnosis, robotic surgery, and drug discovery, among others (Yu et al., 2018). Despite no inclusion of healthcare specific research in the current review, it is logical to assume that AGI systems in healthcare are being considered and may have enormous benefits. Baum (2017) identified 20 active AGI projects with the stated goal being Humanitarian, which may include healthcare, yet was not explicitly stated. As with defence, there would be significant risks associated with AGI systems which have the capacity to make life or death decisions. Moreover, many philosophical discussions focus on healthcare AGI which may seek to optimise achievement of its stated goals via means which provide risk to human life. For example, AGI systems which are tasked with eradicating diseases such as cancer which establish that they can achieve their goal more efficiently by eradicating those in the population who have a genetic predisposition to cancer (Salmon, Hulme et al., 2021) or at the expense of other, longer-term chronic conditions. As such, research on the potential risks associated with the unsafe AGI, or the definition of 'safety' in the healthcare system is necessary future research direction.

The review revealed that a majority of articles do not provide details of the AGI system's specifications or functionality. Rather, most of the articles refer to generic AGI systems without

**Table 2.** Analysis methods identified in the reviewed articles.

Analysis method	Frequency
Fault trees and influence diagrams	1
Risk estimate modelling	1
Philosophical discussion	8
Assessment of current standards and processes	2
Questionnaire/survey	1
Bayesian modelling	1
Diffusion modelling, predator–prey models, and hostility models.	1
Theoretical all-pay contest model	1

describing what capabilities they possess, what goals they may have, and what tasks they will likely perform. It is our view that this is a significant limitation, as it impacts the quality of subsequent risk assessment efforts. Whilst many formal risk assessment methods exist (Dallat, Salmon & Goode 2019), most require at least some description of the tasks being performed and the goals being pursued, and state-of-the-art methods also require a description of the system in which the tasks are being performed (e.g., Dallat et al., 2018; Leveson, 2011; Stanton & Harvey, 2017). Without identifying the specifications of different AGI systems, it is not possible to accurately forecast the range of risks associated with them.

Currently there are few modelling efforts investigating the diffusion of AGI into society (Narain et al., 2019), which was exemplified in the current review by the limited number of modelling approaches identified. Although our understanding of the risk associated with AGI is limited, they have similar characteristics to other risks that involve integration of humans and technology, and modelling techniques exist that may provide meaningful analyses in relation to AGI (Barrett & Baum, 2017). That said, a limitation of the current literature is that there are few studies employing formal scientific analysis methods to identify and assess the risks associated with AGI. Further research exploring the use of modelling approaches such as computational modelling (Salmon et al, 2020), systems analysis (Stanton et al., 2013), and risk assessment methods (Dallat et al., 2018) is recommended.

The review revealed that there is a dearth of peer reviewed literature focused specifically on the risks associated with AGI. Based on the current eligibility criteria, the number of included peer reviewed journal articles that focused on the risks associated with AGI was small. Further, a recent narrative review on catastrophic AGI risks (Sotala & Yampolskiy, 2015) included 314 sources in the references list, of which less than one third were peer reviewed journal articles. Within this subset of articles, many were not directly associated with AI or AGI. While the review by Sotala and Yampolskiy (2015) provided a comprehensive account of the current state of thinking on the risks associated with AGI, it also highlighted that the AGI literature is dominated by non-peer reviewed publications, mainly books and commentaries. It is therefore concluded from this review that there is a critical need for the publication of scientific research focused on the risks associated with AGI within the peer review literature. This is not to diminish the importance of the non-peer reviewed AGI literature, as much of it has shaped our thinking around AGI risk and safety (Bostrom, 2014; Goertzel & Pennachin, 2007). A potential issue with the relatively small amount of peer reviewed articles specifically focused on AGI risk is that policy makers may not take the risks associated with AGI seriously. Global organisations such as the World Health Organisation and the Climate Change Advisory Council base decisions and advice on peer reviewed literature (Alberts et al., 2008; Bornmann, 2011). Furthermore, various legal and regulatory decisions are based on evidence from peer reviewed literature (Bornmann, 2011). As such, peer review is seen as an instrument for ensuring trustworthiness (Cronin, 2005). Although peer review is not without its issues (Holmes et al., 2006), the seriousness of the potential risks from AGI may be underestimated by policy makers. As such there may be a need for more peer reviewed AGI risk research into better engage with key stakeholders such as politicians, advisory groups, and funding bodies.

**Table 3.** Risk categories and definitions identified in the included articles.

Risk category	Definition
AGI removing itself from the control of human owners/managers	The risks associated with containment, confinement, and control in the AGI development phase, and after an AGI has been developed, loss of control of an AGI.
AGIs being given or developing unsafe goals	The risks associated with AGI goal safety, including human attempts at making goals safe, as well as the AGI making its own goals safe during self-improvement.
Development of unsafe AGI	The risks associated with the race to develop the first AGI, including the development of poor quality and unsafe AGI, and heightened political and control issues.
AGIs with poor ethics, morals and values	The risks associated with an AGI without human morals and ethics, with the wrong morals, without the capability of moral reasoning, judgement,
Inadequate management of AGI	The capabilities of current risk management and legal processes in the context of the development of an AGI.
Existential risks	The risks posed generally to humanity as a whole, including the dangers of unfriendly AGI, the suffering of the human race

*Note: Included articles covered one or multiple risk categories*

Lastly, there is no agreed upon single definition of the concept of AGI (Goertzel, 2014). Just as the term AI has many different meanings with the AI community, it appears AGI may suffer the same lack of definitive identity. A finding from the review was the lack of consensus in the terms used within AGI research. An example from the reviewed studies was the reporting of machine intelligence at human level and above. Terms included, AGI, artificial superintelligence (ASI), strong AI, High-level-machine-intelligence (HLHI), superintelligent AI, powerful general artificial intelligence (PAGI), and superintelligent agent. In addition, Goertzel (2014) identified further AGI terms, for example, computational intelligence, natural intelligence, cognitive architecture, and biologically inspired cognitive architecture. A lack of standardised terminology may be confusing what is already a complex research field. Given the potential severe consequences unsafe AGI poses, approaches to make the discipline more accessible to researchers from multiple disciplines is required. Future research could be conducted to standardise a range of AGI terminologies and research fields via consensus of experts through Delphi studies (Linstone & Turoff, 1975).

An important consideration in generalising the findings of the literature is to acknowledge that eligible articles may not have been identified in the literature search. However, the search strategy included three academic databases and included comprehensive Boolean search terms and filters. In addition, relevant contemporary literature might be in formats excluded in this review, such as conference proceedings, grey literature, reviews, arXiv articles, journal editorials and books. For example, it is acknowledged that publishing novel research through peer reviewed conference proceedings is standard practice in the computer science.

## Conclusion

The current systematic review was conducted to investigate the extant peer reviewed literature focused on the risks associated with AGI. Data extracted from the eligible articles included, the type of analysis methods used, risks associated with AGI, and recommended risk controls/risk management strategies. From the small number of eligible articles, a broad range of risks were identified including, AGI removing itself from the control of human owners/managers, AGIs being given or developing unsafe goals, development of unsafe AGI, AGIs with poor ethics, morals and

values, inadequate management of AGI, and existential risks. However, issues with the current state of peer reviewed AGI risk literature emerged. First, there was a scarcity of modelling techniques applied to investigate risks associated with AGI. Second, there was a limited number of studies that focused on the AGI risks in specific domains. Third, the lack of information regarding the AGI systems considered in terms of specifications, goals and tasks raises questions about the validity and comprehensiveness of the risks identified. Fourth, there was a limited amount of peer reviewed literature on the risks of AGI. Finally, there is a lack of consensus on the terminology used within AGI research. It is concluded that there is a critical need to address the multiple issues identified in the current review. Given that the fate of humanity may be at stake with the development of unsafe AGI, it is essential that we have reliable, valid, and rigorous research to guide safe AGI design, implementation and management.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Australian Research Council [DP200100399].

## ORCID

Scott McLean  <http://orcid.org/0000-0002-7269-5847>

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

## References

- Alberts, B., Hanson, B., & Kelner, K. L. (2008). Reviewing peer review. *Science*, 321(5885), 15. <https://doi.org/10.1126/science.1162115>
- Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking inside the box: Controlling and using an oracle AI. *Minds and Machines*, 22(4), 299–324. <https://doi.org/10.1007/s11023-012-9282-2>
- Barrett, A. M., & Baum, S. D. (2017). A model of pathways to artificial superintelligence catastrophe for risk and decision analysis [Article]. *Journal of Experimental and Theoretical Artificial Intelligence*, 29(2), 397–414. <https://doi.org/10.1080/0952813X.2016.1186228>
- Baum, S. (2017). A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, 17–11, Global Catastrophic Risk Institute.
- Baum, S. D., Barrett, A. M., & Yampolskiy, R. V. (2017). Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica-Journal of Computing and Informatics*, 41(4), 419–427. <http://www.informatica.si/index.php/informatica/article/view/1812>
- Baum, S. D., Goertzel, B., & Goertzel, T. G. (2011). How long until human-level AI? Results from an expert assessment. *Technological Forecasting Social Change*, 78(1), 185–195. <https://doi.org/10.1016/j.techfore.2010.09.006>
- Bentley, P. (2018). *The Three Laws of Artificial Intelligence: Dispelling Common Myths* European Parliamentary Research Service. European Parliamentary Research Service.
- Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science Technology*, 45(1), 197–245. <https://doi.org/10.1002/aris.2011.1440450112>
- Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolutionary Technology*, 9. <https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press Inc.
- Boyles, R. J. M. (2018). A case for machine ethics in modeling human-level intelligent agents [Article]. *Kritike*, 12(1), 182–200. <https://doi.org/10.25138/12.1.a9>
- Bradley, P. (2020). Risk management standards and the active management of malicious intent in artificial superintelligence [Article]. *AI & Society*, 35(2), 319–328. <https://doi.org/10.1007/s00146-019-00890-2>
- Bringsjord, S., Bringsjord, A., & Bello, P. (2012). *Belief in the singularity is fideistic (Singularity Hypotheses (pp. 395-412))*. Springer.
- Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental and Theoretical Artificial Intelligence*, 26(3), 355–372. <https://doi.org/10.1080/0952813X.2014.895108>

- Chalmers, D. (2009). The singularity: A philosophical analysis. *Science Fiction and Philosophy: From Time Travel to Superintelligence*, 171-224. <https://www.wiley.com/en-au/Science+Fiction+and+Philosophy%3A+From+Time+Travel+to+Superintelligence-p-9781444327908>
- Chen, S. Y., & Lee, C. (2019). Perceptions of the impact of high-level-machine-intelligence from University students in Taiwan: The case for human professions, autonomous vehicles, and smart homes. *Sustainability*, 11 (21), 6133. Article 6133. <https://doi.org/10.3390/su11216133>
- Cronin, B. (2005). *The hand of science: Academic writing and its rewards*. Scarecrow Press.
- 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>
- Dallat, C., Salmon, P. M., & Goode, N. (2019). Risky systems versus risky people: To what extent do risk assessment methods consider the systems approach to accident causation? A review of the literature. *Safety Science*, 119, 266-279.
- Firt, E. (2020). The missing G. AI & SOCIETY, 1-13.
- Garis, H. D. (2005). *The artelect war: Cosmists vs Terrans*. ETC Publications.
- Goertzel, B. (2006). *The hidden pattern*. Brown Walker Press.
- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1-48. <https://doi.org/10.2478/jagi-2014-0001>
- Goertzel, B., & Pennachin, C. (2007). *Artificial general intelligence* (Vol. 2). Springer.
- Goertzel, B., & Pitt, J. (2014). Nine ways to bias open-source artificial general intelligence toward friendliness. In Russell Blackford, & Damien Broderick (Eds.), *Intelligence unbound* (pp. 61-89). Wiley.
- Holmes, D., Murray, S. J., Perron, A., & Rail, G. (2006). Deconstructing the evidence-based discourse in health sciences: Truth, power and fascism. *International Journal of Evidence-Based Healthcare*, 4(3), 180-186. <https://doi.org/10.1111/j.1479-6988.2006.00041.x>
- 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>
- Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. Penguin.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Legg, S., & Hutter, M. (2006). A formal measure of machine intelligence. arXiv preprint cs/0605024.
- Leveson, N. G. (2011). Applying systems thinking to analyze and learn from events. *Safety Science*, 49(1), 55-64.
- Linstone, H. A., & Turoff, M. (1975). *The delphi method*. Addison-Wesley Reading, MA.
- Miller, J. D. (2019). When two existential risks are better than one. *Foresight*, 21(1), 130-137. <https://doi.org/10.1108/FS-04-2018-0038>
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, 6(7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Müller, V. C., & Bostrom, N. (2016). *Future progress in artificial intelligence: A survey of expert opinion (Fundamental issues of artificial intelligence)* (pp. 555-572). Springer.
- Narain, K., Swami, A., Srivastava, A., & Swami, S. (2019). Evolution and control of artificial superintelligence (ASI): A management perspective. *Journal of Advances in Management Research*, 16(5), 698-714. <https://doi.org/10.1108/JAMR-01-2019-0006>
- Naudé, W., & Dimitri, N. (2020). The race for an artificial general intelligence: Implications for public policy [Article]. *AI & Society*, 35(2), 367-379. <https://doi.org/10.1007/s00146-019-00887-x>
- Nindler, R. (2019). The United Nation's capability to manage existential risks with a focus on artificial intelligence. *International Community Law Review*, 21(1), 5-34. <https://doi.org/10.1163/18719732-12341388>
- Pueyo, S. (2018). Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, 197, 1731-1736. <https://doi.org/10.1016/j.jclepro.2016.12.138>
- Salmon, P. M., Carden, T., & Hancock, P. (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.
- Salmon, P. M., Hulme, A., Walker, G. H., Waterson, P., Berber, E., & Stanton, N. A. (2020). The big picture on accident causation: A review, synthesis and meta-analysis of AcciMap studies. *Safety Science*, 126, 104650. <https://doi.org/10.1016/j.ssci.2020.104650>
- Salmon, P. M., Read, G. J., Thompson, J., McLean, S., & McClure, R. (2020). Computational modelling and systems ergonomics: a system dynamics model of drink driving-related trauma prevention. *Ergonomics*, 63(8), 965-980. <https://doi.org/10.1080/00140139.2020.1745268>
- Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4). <http://www.informatica.si/index.php/informatica/article/view/1877/1098>
- Sotala, K., & Yampolskiy, R. V. (2015). Responses to catastrophic AGI risk: A survey. *Physica Scripta*, 90(1). 1-33. Article 018001. <https://doi.org/10.1088/0031-8949/90/1/018001>



- Stanton, N., Salmon, P. M., & Rafferty, L. A. (2013). *Human factors methods: A practical guide for engineering and design*. Ashgate Publishing, Ltd.
- Stanton, N. A., Eriksson, A., Banks, V. A., & Hancock, P. A. (2020). Turing in the driver's seat: Can people distinguish between automated and manually driven vehicles? *Human Factors Ergonomics in Manufacturing Service Industries*, 30(6), 418–425. <https://doi.org/10.1002/hfm.20864>
- 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.
- Stanton, N. A., Eriksson, A., Banks, V. A., & Hancock, P. A. (2020). Turing in the driver's seat: Can people distinguish between automated and manually driven vehicles? *Human Factors and Ergonomics in Manufacturing & Service Industries*, 30(6), 418–425.
- Tegmark, M. (2017). *Being human in the age of artificial intelligence*. Vintage Books.
- Torres, P. (2019). The possibility and risks of artificial general intelligence. *Bulletin of the Atomic Scientists*, 75(3), 105–108. <https://doi.org/10.1080/00963402.2019.1604873>
- Yampolskiy, R. V. (2012). Leakproofing singularity-artificial intelligence confinement problem. *Journal of Consciousness Studies*, 19(1-2), 194–214. <https://www.ingentaconnect.com/contentone/imp/jcs/2012/00000019/f0020001/art00014>
- Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10), 719–731. <https://doi.org/10.1038/s41551-018-0305-z>