AI & SOCIETY
https://doi.org/10.1007/s00146-019-00890-2

ORIGINAL PAPER

Risk management standards and the active management of malicious intent in artificial superintelligence

Patrick Bradley[1]

Received: 14 August 2018 / Accepted: 2 April 2019

Abstract
The likely near future creation of artificial superintelligence carries significant risks to humanity. These risks are difficult to conceptualise and quantify, but malicious use of existing artificial intelligence by criminals and state actors is already occurring and poses risks to digital security, physical security and integrity of political systems. These risks will increase as artificial intelligence moves closer to superintelligence. While there is little research on risk management tools used in artificial intelligence development, the current global standard for risk management, ISO 31000:2018, is likely used extensively by developers of artificial intelligence technologies. This paper argues that risk management has a common set of vulnerabilities when applied to artificial superintelligence which cannot be resolved within the existing framework and alternative approaches must be developed. Some vulnerabilities are similar to issues posed by malicious threat actors such as professional criminals and terrorists. Like these malicious actors, artificial superintelligence will be capable of rendering mitigation ineffective by working against countermeasures or attacking in ways not anticipated by the risk management process. Criminal threat management recognises this vulnerability and seeks to guide and block the intent of malicious threat actors as an alternative to risk management. An artificial intelligence treachery threat model that acknowledges the failings of risk management and leverages the concepts of criminal threat management and artificial stupidity is proposed. This model identifies emergent malicious behaviour and allows intervention against negative outcomes at the moment of artificial intelligence's greatest vulnerability.

Keywords Risk · ISO 31000 · Threat · Existential · Superintelligence · Artificial intelligence · Artificial stupidity · Criminal threat management

1 Introduction

Many experts think that machines many times more intelligent than humans will exist by 2075 and that some form of superintelligent machines will exist within the next 25–50 years (Bostrom 2014; Brundage et al. 2018; Meek et al. 2016). There are many paths to creating superintelligence, and numerous private organisations and governments are working on developing increasingly powerful artificial intelligence (AI) The numerous paths to creating superintelligence and the massive strategic advantage it would give to any organisation or government not only makes the future creation of an artificial superintelligence (ASI) inevitable it amplifies the risks as the strategic pressure to create an ASI is likely to relegate safety to a low priority.

There is, of course, significant debate about the likelihood of superintelligence occurring and debate about the expected timelines (Baum et al. 2011). For this paper, the assumption is that superintelligence is an achievable reality that will happen within the time frame of 25 to 50 years with increasingly powerful AI leading to the 50-year upper limit.

Many prominent figures in science and technology such as Stephen Hawking (BBC 2014) and Elon Musk (Sydney Morning Herald 2017) hold the opinion that superintelligence poses the greatest risk to humanity of any of the threats we face today. It is a threat that far exceeds the risks of climate change, overpopulation, and nuclear war.

In his own words, Musk says *"With artificial intelligence, we are summoning the demon. You know all those stories where there's the guy with the pentagram and the holy water and he's like, yeah, he's sure he can control the demon? Doesn't work out"* (Dowd 2017).

✉ Patrick Bradley
patrick@patrick.ninja

[1] Thrive Plus Consulting, Level 35, Tower One, 100 Barangaroo Avenue, Sydney, NSW 2000, Australia

Published online: 11 April 2019

Currently available AI is already altering the types of risks individuals, organisations and states are exposed to and the malicious use of AI poses significant and poorly understood risks to digital security, physical security and the integrity of Western democratic systems. The malicious use of AI by humans is already well established and the potential for major societal impacts is likely to significantly increase as AIs become more sophisticated (Brundage et al. 2018).

The implications of a malicious ASI that is many millions of times more capable than current single purpose AIs are hard to conceptualise (Brundage et al. 2018) Identifying and combating this threat is a new frontier for risk professionals and this paper argues that the risk assessment portion of ISO 31000, the current globally accepted model for risk management and the model on which most risk management is based, is not fit for this purpose. Risk managers must understand that the current model will not work in the face of the grave existential risks posed by ASI and a thorough reconceptualisation of how to manage these types of risks is necessary.

To address the failings of ISO 31000 alternatives to risk management will be discussed and a preliminary model to manage the threat of malicious ASI will be proposed.

The intended audience of this paper is risk professionals and AI developers[1] with a view to bridging knowledge and objectives gaps between the two groups with an end goal of safer AI. As such, subjects that may be seen as common knowledge to AI developers will be discussed in some detail and vice versa for risk professionals.

2 Current use of risk management in AI development

There exists numerous books, articles, blogs, analysis and opinions on the risks posed by AI (BBC 2014; Bailey 2017; Bostrom 2014; Dowd 2017; Future of Life Institute 2017; Goertzel 2015). However, there is minimal published research on how risk is currently being managed and even less practical guidance in the form of tools or standards.

For example, Google and Facebook outline their principles behind making AI safe. But beyond clarifying what they will not develop[2] there is very little detail about how safety is being practically achieved (Google 2018; Facebook 2019). Given the commercial and strategic value in AI it is unsurprising that organisations would reveal very little besides efforts to manage public perceptions that they are operating safely and ethically.

Open AI, a non-profit research organisation whose mission is to "build safe AI" is another example of the lack of practical guidelines (OpenAI 2019a, b). OpenAI is not bound by commercial or strategic interests and seeks to keep AI research open so that safety can be maximised, and to ensure the benefits of AI are not limited to large organisations.

OpenAI supports developers with tools, software and a community of peers. OpenAI publish software tools that could, among other uses, have a safety application such as Gym, a toolkit for developing reinforcement learning (OpenAI 2019a, b). However, even OpenAI, one of the premier sources of openly available AI tools does not clearly state how a developer should manage the risks of their work.

A study of current risk management in AI development is beyond the scope of this paper; therefore, instead a parallel will be drawn to general information technology projects. Most AI development would be categorised as an IT project by virtue of its primary focus on technology, its definitive beginning and once the objective is met the work on developing that objective will stop or move to another objective.

Identifying AI development as an IT project is consistent[3] with the definition in the project managers book of knowledge (PMBOK) which is a de-facto international standard for project management,[4] including IT projects[5] (Jamali and Oveisi 2016).

Almost every government institution or organisation has a project management framework of some sort to manage budgets, schedules, requirements and risk. There are countless project methodologies in use such as Agile, Waterfall and Critical Path Method (Cohen 2017).

[1] The generic term AI developer will be used throughout this paper to as a catch all reference to anyone (or any organisation) involved in the development, research and implementation of artificial intelligence technologies.

[2] For example, Google has stated it will not participate in developing weapons systems (Google 2018).

[3] PMBOK defines a project as: "…a temporary endeavour undertaken to create a unique product, service, or result. The temporary nature of projects indicates that a project has a definite beginning and end. The end is reached when the project's objectives have been achieved or when the project is terminated because its objectives will not or cannot be met, or when the need for the project no longer exists." (PMBOK® Guide—Sixth Edition 2017).

[4] PRINCE2 is another standard used in the UK, Australia and many European countries (Karaman and Kurt 2015). It has many commonalities with PMBOK and risk management is a common feature in almost all project management methodologies (Jamali and Oveisi 2016). A comparison of project management standards is beyond the scope of this paper.

[5] Risk management in IT projects is a distinct activity from IT security management which is covered by ISO/IEC 27005:2018 Information technology - Security techniques - Information security risk management. While they share numerous commonalities ISO/IEC 27005:2018 is specifically focussed on information security management and is not a general risk methodology like ISO 31000.

Whereas ISO 31000 provides an overview of how to manage risk in any activity, PMBOK provides detailed guidance and tools on how to specifically manage project risk.[6] The methodology used in PMBOK is a direct derivative of ISO 31000. The process and terminology in PMBOK, while having some project specific tools and language to make it more project relevant, is the same as ISO 31000, and therefore, has the same points of failure (PMBOK® Guide—Sixth Edition 2017). Risk management being a central feature of project management is common across the various common project methodologies (see Table 1).

On the basis of AI development having a strong alignment with project management, the almost universal use of project management methodologies (such as PMBOK) to manage projects and the use of risk management tools in these project management methodologies it can be asserted that there is a high likelihood that most AI development is actively using risk management tools. Therefore, the issues identified with risk management in this paper are likely to apply to most AI development.

This subject area needs further research. The approaches to risk management in AI development should be studied to ascertain how risk is being practically managed and assess the theoretical effectiveness of those approaches if they differ significantly to ISO 31000.

3 Existential risks

"Existential risks have a cluster of features that make ordinary risk management ineffective" (Bostrom 2002).

Bostrom developed the idea of Existential Risks in his 2002 paper "Existential Risks: Analysing Human Extinction Scenarios and Related Hazards". In this paper, he described the features of Existential Risks as *"…where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential"*. These would be events such as the deliberate misuse of nanotechnology, total nuclear war, simulation shutdown and badly programmed superintelligence (Bostrom 2002).

While Bostrom recognised that risk management was ineffective for dealing with existential risks he did not specify why it was ineffective at an implementation level. This paper will develop his ideas as they relate to the risks of superintelligence using current risk management standards.

Table 1 Risk management in project management

Project management methodology	Is risk management a significant part of the methodology?
PMBOK	Yes, as one of the 10 knowledge areas (PMBOK® Guide—Sixth Edition 2017)
PRINCE2	Yes, as the management of risk (MoR) method (AXELOS Limited 2014)
Agile	Yes, Agile has tools specifically to deal with risk such as the risk burndown chart (Moran 2014)

4 Superintelligence

A discussion of what constitutes Superintelligence is beyond the scope of this paper. For this paper, Superintelligence will be broadly defined using Bostrom's description of a non-human (artificial) intelligence that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills (Bostrom 2006). The term AI will be used to reference a pre-superintelligent artificial intelligence and ASI will be used to denote artificial superintelligence.

5 Defining the common model for risk management

ISO 31000:2018, the current global standard for risk management, is a quantitative and qualitative model of risk that defines risk as "the effect of uncertainty on objectives". (International Organisation for Standardisation 2018) ISO 31000 falls into a broad category of risk management methodologies often loosely defined as operational risk management (Raz and Hillson 2005).

ISO 31000 is not only used to minimise negative outcomes but also to maximise positive opportunities. The focus of this paper will be how the methodology works to minimise negative outcomes. For the purposes of this paper, the objective is the survival and continued positive growth of the human race.

5.1 ISO 31000 process

This paper will focus on the risk assessment portion of ISO 31000 (section 6.4 of the standard.) contained within section 6.4 of ISO 31000 is risk identification (section 6.4.2), an assessment of likelihood (section 6.4.3) and consequence (section 6.4.3); these will be the primary areas of analysis.

[6] Risk management is a central feature of project management and PMBOK defines project risk management as: "Project Risk Management includes the processes of conducting risk management planning, identification, analysis, response planning, and controlling risk on a project. The objectives of project risk management are to increase the likelihood and impact of positive events and decrease the likelihood and impact of negative events in the project." (PMBOK® Guide—Sixth Edition 2017).

Risk assessment is not the only focus of ISO 31000. It provides guidelines around many other dimensions of risk management but the risk assessment (6.4) component is fundamental to the effectiveness of the entire standard and the effectiveness of the overall process is compromised if the risk assessment function does not work as intended.

Risk, as described in ISO 31000 is a function of likelihood and consequence, it can be simplified as shown below.[7,8,9]

$R_x = f(C_x \text{ and } L_{cx})$

x, the identified risk type; R_x, risk level of identified risk x; C_x, consequence of risk x; L_{cx}, likelihood of consequence[10] x.

5.2 Risk management has a common architecture

While there exists a diverse vocabulary of terms, variations in process and different implementation strategies, most non-ISO 31000 approaches to risk management are built around the same basic architecture (Raz and Hillson 2005; International Organisation for Standardisation 2018; Hudsal 2015).

The common failure points discussed in this paper appear, sometimes with different names, in many other risk standards besides ISO 31000. As such, the term "risk management" will be used throughout this paper to signify ISO 31000 and any other risk processes with the same common features (Table 2).

6 Issues with risk management architecture and superintelligence

Within normal risk management, bias is an issue that can negatively impact the effectiveness of a risk plan (Taylor and Blaskovich 2011; Heemstra and Kusters 2003; Harding 2016a). Bias in the risk process is amplified when dealing with ASIs due to one of our most common personality traits, illusionary superiority.

The Dunning–Kruger effect is a bias whereby people of low ability suffer from illusionary superiority which leads them to consider their own cognitive ability as superior (Kruger and Dunning 1999). The Dunning–Kruger effect is in a similar category of human positive illusions to the Overconfidence Effect (Pallier et al. 2002; Pennycook et al. 2017) and Illusionary Superiority (Hoorens 1993). While these three biases are all slightly different, they all lead us to underestimate the abilities of others and overestimate our own abilities, especially when dealing with issues that outwardly appear simple but are in fact complex.

Given the pervasive nature of human biases stemming from positive illusions, it is entirely reasonable to expect that many people involved in developing ASI will approach the problem of risk with those biases. Given that an ASI is likely to have access to the entirety of human knowledge it is also reasonable to expect that it would exploit these biases in its early stages of development to meet its objectives. This further reduces the likelihood that risk management could effectively manage the risks of the ASI as it all serves to erode fundamental aspects of the risk management process: risk identification, likelihood estimation and consequence estimation.

6.1 Risk identification and anthropomorphic bias

Risk management requires accurate risk identification (International Organisation for Standardisation 2018; Raz and Hillson 2005). There is simply no data on what an ASI may do, and any attempt at accurate risk identification is likely to be too broad to be useful and too tarnished with anthropomorphic bias. Studies have shown that even in simulations where the participants clearly understand an AI is not human, they will assign it human qualities like empathy and reasoning (Barrett 1996). As such, we simply could not reliably identify risks without interjecting human bias into the process. An ASI would likely recognise and exploit this bias.

A competent risk manager often deals with unclear risk identification using boilerplate risks (these are risks that are common to the area being managed) and iteratively fine-tuning the risk plan from there by looking at historical data, talking to experts and working with stakeholders.

A basic scenario of an ASI attack could be along the lines of gaining network access, propagating across networks and taking control of resources. A risk manager could build a good risk management plan around this sort of event sequence, however, it is unlikely that an ASI would follow the human playbook and would attack in much more subtle and unpredictable ways. For this reason, dealing with a lack of clearly identified risks using boilerplate risks is unlikely to be effective.

6.2 Likelihood and consequence in risk analysis

There exists no data about what specific risks an ASI poses and the consequences of those risks. While risk management is sometimes driven by conjecture, especially with very low likelihood events, there is almost always historical

[7] Risk score is a function of consequence and likelihood, not a pure mathematical summation (Lowder 2010).

[8] Risk is sometimes calculated using the formula: Risk = function of (Threats, Vulnerabilities and Consequence) In this case Threats and Vulnerabilities effectively equal likelihood (Lowder 2010).

[9] The simplified formula for calculating risk isn't explicitly stated in ISO 31000, this formula is a derivative of the standard.

[10] The ISO 31000: 2009 version defined likelihood as the likelihood of the consequence occurring. The 2018 version has a broader definition to include the likelihood of the event occurring. Either interpretation works for the purposes of this paper.

Table 2 Examples of risk model commonalities across methodologies

ISO 31000: 2018 (International Organisation for Standardisation 2018)	IT infrastructure library service lifecycle (Office of Goverment Commerce 2007)	Failure mode effect analysis (GE Power Systems University 2001) and (Pande et al. 2007)	Project Management Book of Knowledge (PMBOK® Guide—Sixth Edition 2017)	Risk management for security professionals (Roper 1999)
Risk identification	Threat	Failure mode	Identify risks	Potential undesirable event
Likelihood	Probability of threat and vulnerability	Likelihood of occurrence	Probability of occurrence	Threat and Vulnerability rating
Consequence	Impact	Degree of severity	Impact on objective	Impact rating
Risk level	Risk level	Risk priority	Risk rating	Overall risk
Risk treatment	Risk reduction measures	Controls	Risk response	Countermeasures

data. Even with incredibly low likelihood/high consequence events such as asteroid strikes there are historical records. There is no data on what an ASI would or could do, and as such, the corresponding likelihood of a consequence occurring are also unknown.

The lack of likelihood data is also a problem in industries such as nuclear power where some types of predicted disasters have never occurred. To deal with this, they create probability models by deconstructing the known failure mode and analysing the failure probability of individual mechanical components and building a risk model from that (Hubbard 2015). However, this approach will not work with ASI's because the identified risk is unknown.

The lack of likelihood data and well-understood consequences severely impedes the ability of the risk process to be effective at managing risk. ISO 31000 acknowledges the difficulty of quantifying highly uncertain, high consequence events and recommends a combination of risk techniques to get better insight (International Organisation for Standardisation 2018). However, a combination of techniques will still draw from the same flawed set of tools and assumptions about risk.

6.3 Common mode failure

Risk management tends to deal with single risks and follow them through from identification and analysis to mitigation. However, many failure events occur in a common mode whereby a single failure can either create further failures in the same process; or impact other processes in unexpected ways (Wang and Roush 2000). For example, a cable failure in a chemical plant may damage backup equipment unrelated to the cable, and therefore, massively change the risk profile of a process unrelated to the original event. These common mode risks are not effectively dealt with using ISO 31000 risk management. There are tools to manage these types of highly uncertain low probability risks such as Monte Carlo Analysis, but they do not form part of the standard risk management architecture.

As with the other failure points, an ASI is likely to identify that common mode risks are difficult to manage and actively exploit them.

7 Superintelligence is a dynamic malicious threat actor

"By far the greatest danger of Artificial Intelligence is that people conclude too early that they understand it" (Yudkowsky 2008)

Risk management deals best with static risks that have clear pathways to identification, well-defined likelihoods, clear consequences and feasible ways to mitigate the risk. However, despite the implementation of a competent risk management plan with all the associated tactics and procedures, competent malicious threat actors such as professional criminals and determined terrorists can overcome risk mitigation strategies and launch a successful attack (Harding 2014, 2016a, b).

The use of the word "malicious" is one of convenience, it is unlikely that Superintelligence would hold any malicious feelings towards humans. Malicious is generally defined as "the intention to cause harm". An ASI is likely to have the intention to do whatever it wants with no value judgements about harmful outcomes. As Eliezer Yudkowsky explains *"The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else"* (Yudkowsky 2008). However, if superintelligence was to try and destroy all humans, it would likely seem very malicious to us.

Bostrom identifies the start point of this malicious behaviour as being a "treacherous turn". The ASI would know that its release from a controlled "sandbox" environment was contingent on it being cooperative and friendly. Once it has passed this flawed test of safety and is released into an uncontrollable and infinitely scalable environment, it is free to do as it pleases (Bostrom 2014).

There exists a range of theoretical approaches to solve the AI Control Problem. Approaches such as redundant safety measures, tripwires, adversarial architectures, formalising suffering and benign test environments, amongst others, have been proposed. They all are, to various degrees, essentially a risk mitigation strategy for the perceived risks surrounding ASIs as they assume a known set of outcomes from emergent malicious behaviour. They are, therefore, likely to suffer many of the same issues the risk management process suffers; poorly established risk types, unknown probabilities, undefined consequences and inherent human bias (Bostrom 2006, 2014; Bailey 2017; Yampolskiy 2012; Baumann 2017; Chalmers 2010).

7.1 Criminal threat management and leveraging artificial stupidity to identify the treacherous turn

Risk management has many critics both of its architecture and the failure to properly implement risk management plans (Viljoen and Musvoto 2013; Kimball 2000; Sabato 2009; Power 2009; Harding 2016a, b). Within the field of security risk management, the criticisms occur for similar reasons to the problems encountered with an ASI; highly competent and motivated malicious threat actors such as professional criminals and terrorists will identify how risks are being managed and attack in unexpected ways to achieve their goal.

Harding (2016a, b) is a strong critic of the use of risk management in security and goes so far as calling it a *"dangerously overrated and broken paradigm"*. As an alternative to the failings of risks management he proposes an approach when dealing with malicious threat actors called Criminal Threat Management (Harding 2014, 2016a, b).

Harding differentiates criminal threat management from risk management because it focuses on the malicious threat actor, whereas risk management primarily focuses on minimising detrimental outcome events (Harding 2014). Harding's criminal threat management concept is a useful one for managing ASIs as its baseline assumption is that determined malicious threat actors will overcome risk management-based countermeasures through skill, capability and adaptive dynamic tactics.

Several researchers (Harding 2014; Chun and Lee 2013; Smith and Louis 2010) have identified that the methodology professional criminals or terrorists utilise is fairly consistent. There are two main relevant components of the criminal threat management cycle that are very relevant to ASIs; developing the intent and acting on that intent (intent to actualisation). The relevance of these two components is that they frame the model for non-malicious ASI around the idea of stopping the development of malicious intent and blocking or guiding the actualisation of malicious intent.

Bostrom proposes some ways to guide or stop the development of malicious intent. Solutions such as teaching an ASI human values and for the ASI to compare its goals and behaviour to those "learnt" human values is one solution. Another option is having the ASI watch human behaviour and determine normative standards for human desires. By learning from observation and self-correcting the ASI could learn over time and even change its standards as human standards change (Bostrom 2014).

Bostrom's idea of observing or teaching normative human values is consistent with Harding's criminal threat management cycle concept of guiding and countering the development of intent. The criminal threat management cycle coupled with the failings within the risk management architecture when applied to ASIs would suggest that any attempt to guide or counter intent needs to occur before there is any real capability to inflict harm as the likelihood of the ASI successfully deceiving us increases as its capabilities increase.

Therefore, any normative human standards or baseline acceptable human behaviour needs to be instilled in the ASI very early in the development cycle *before* there is any chance of awareness and well before it develops any meaningful intelligent capabilities.

If "guiding intent" can be successfully implemented in the ASI there would also need to be some method to warn humans if the ASI was considering deviating from normative human standards. Coming back to professional criminals and terrorists, there exist a number of common factors that occur in the "intent to actualisation" phase of an attack. These are things such as target identification, intelligence gathering, planning, preparation and dry-runs or reconnaissance (Harding 2014).

A similar set of "trigger" behaviours could be established for an ASI and if it appeared that it was deviating from normal standards of behaviour whatever mechanism was being used to enforce normative behaviour could be re-applied to block or disrupt the non-normative behaviour. Trigger behaviours will be discussed in detail in a subsequent section.

8 Issues with the criminal threat management approach

There are issues with applying the criminal threat management approach to an ASI. The issue of anthropomorphic bias is not fully resolved as the intervention options are still largely based upon human logic and it is entirely possible that a malicious ASI would approach the barriers of non-normative behaviour blocks in a way that is completely outside the range of human thinking. However, because there exist no trial and error approach the baseline assumption of

this approach is that the ASI is a malicious threat actor and safety must be built in long before there is any possibility of human level intelligence.

Another issue with the approach of guiding and countering intent is that it simply may be too cumbersome and the speed with which an ASI evolves intent and capabilities may overwhelm any human attempts at intervention with speed. This is yet another reason why safety protocols must be built into the basic architecture of any AI/ASI as a first step.

Finally, as Yudkowsky (2015) points out "value is fragile". Human values have commonly accepted meanings to humans and outside of a human cultural context they may evoke very different meanings (Yudkowsky 2015; Goertzel 2015).

Bostrom (2014) characterises the achievement of benign goals through deleterious unforeseen methods as being "Perverse instantiation" (Bostrom 2014). This can also be characterised as "reward hacking[11]" (Amodei et al. 2016) Any attempt to leverage criminal threat management concepts will face very significant challenges in developing non-fragile human values that are not susceptible to reward hacking

9 Treachery threat management model

As previously discussed, Bostrom identifies the start point of ASI malicious behaviour as being a "treacherous turn" (Bostrom 2014). It is possible that an ASI, capable of strategic planning with time to wait could hide its malicious intentions and capabilities until it has sufficient freedom, abilities and resources to prosecute those malicious intentions (Bostrom 2014; Grace 2015).

The fundamental problem with the Bostrom's treacherous turn argument is that it suggests that there may be no evidence to show that an ASI has undergone a treacherous turn, or vice versa. As noted by Danaher (2014) because the treacherous turn is effectively unfalsifiable, it allows for a possibility that there currently exists a malicious ASI that is just concealing it intentions and powers from us (Danaher 2014; Grace 2015).

While the unfalsifiable treacherous turn may be in principle possible it is a poor basis for developing practical tools for developers to create and manage safe AI. If a falsifiable treacherous turn was a possibility, there is an opportunity to leverage the indicators of treachery and divert the ASI from treachery in that moment of vulnerability. The treacherous turn is the ASI's moment of greatest vulnerability because it is likely that the ASI would be blocked by humans from any further development once it has revealed its malicious intentions.

Trazzi and Yampolskiy's concept of artificial stupidity provides a starting point for triggers that may indicate treachery and unifying the concepts of artificial stupidity and the treacherous turn gives us a basic model to assess treachery while an ASI is still vulnerable (Trazzi and Yampolskiy 2018; Goertzel 2015).

Trazzi and Yampolskiy's model of artificial stupidity poses a number of constraints that could be applied to AI's to limit their capability and stop them transitioning to maliciousness or limit the capability of an emergent ASI. Some examples are: limiting computing hardware power (storage capacity, memory, processing power, clock speed, etc.), limiting software capabilities by removing any capability for self-improvement and introducing programmed bias into the software that favours friendliness (Trazzi and Yampolskiy 2018; Goertzel and Pitt 2012).

Cognitive Stenography could be another significant input into the model (Goertzel 2014). Goertzel proposes a model of cognitive stenography for intelligent agents that becomes increasingly complex as the intelligent agent becomes more cognitively powerful. Within this model is the concept of "stuckness" whereby cognitive processes within the intelligent agent help each other when they become stuck (Goertzel 2014) An input into the model could be a measure of "stuckness", unexpected resolution of "stuckness" or synergy between cognitive processes. Unanticipated changes in any of these dimensions could potentially be leveraged to identify treachery or emergence of intelligence beyond the parameters of safety.

9.1 Artificial intelligence treachery threat model

Based up on the early intervention concepts of criminal threat management, the theoretical failure of predictive risk management and the ideas of artificial stupidity; a preliminary model[12] to manage the threat of a treacherous turn and intervene at the moment of greatest vulnerability is proposed.

The model would require real-time monitoring of factors that either constrain the ability to be treacherous or enforce artificial stupidity. If these factors change in a statistically significant way the ASI could either be re-directed to normative behaviour enforcement, have its capabilities blocked or just powered down. Due to the volume of data and the need for a fast response, this model of enforcement would ideally be managed by a separate AI measuring the potentially treacherous ASI in parallel to a human.

[11] Not all risks from ASIs may be due to malicious behaviour. Reward hacking may lead an ASI to unintentionally cause harm by aggressively pursuing a goal in an unexpected way with negative consequences.

[12] This model could either be used on an AI to detect emergent ASI capabilities or detect abnormal behaviour in an ASI. It is not specific to either definition.

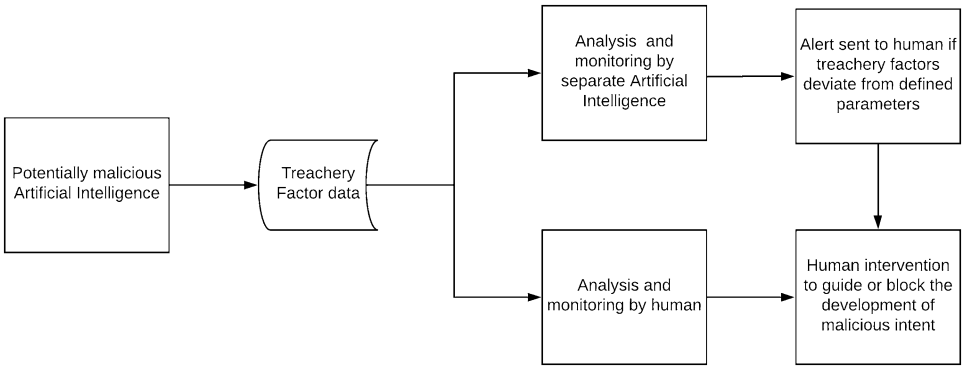

Fig. 1 Artificial intelligence treachery threat model

For example; three factors that may indicate treacherous behaviour that would be relatively easy to measure in real time such as power usage, processer utilisation and clock speed could be tracked using embedded sensors in the ASI's hardware. These sensors could track and log their relevant data point to a database that could, via an artificially intelligent monitoring system, alert a human if deviation from normal parameters was detected. The human could then perform a manual intervention on the ASI that is behaving abnormally.

In this example, the embedded sensors would need to be totally isolated from the ASI under monitoring, as would the AI that is monitoring its performance. The ideas proposed by Goertzel (2014) on cognitive stenography would need to be included in the model to ensure that there is no cognitive collusion between the ASI that is being monitored, the embedded sensors, and the AI that is managing and reporting on performance. This is also why it is critical that a human remains in the monitoring and decision chain. Given the potential power a human may be able to wield with an emergent superintelligence there would also need to be well established safety protocols to monitor the behaviour of the human (Fig. 1).

This type of real-time monitoring where AIs work with humans to manage IT performance factors is already available to consumers and the technology will only improve as the datasets available for machine learning grow (Cisco 2019).

With enough data, even rudimentary statistical tools such as a Shewhart chart[13] would be able to accurately detect variation within a dataset that is abnormal. Much more advanced mathematical models are available and useable with this type of data and deviation from normal process performance is a well understood discipline of statistics (Wood 2002).

This approach of using data-driven predictive models to predict malicious behaviour is not new and current models using machine learning technologies have a high degree of theoretical accuracy in detecting terrorist behaviour (Salih et al. 2017; RAND National Security Research Division 2005; Li et al. 2018; Ding et al. 2017; Schneider et al. 2011). With accessibility to more data for machine learning and iterative improvements, these terrorist focussed models are likely to improve over time and the learnings from these models are likely to be useful inputs into an AI treachery model.

10 Discussions and conclusions

There exists a common architecture within risk management. When applied to an ASI this architecture suffers from fatal flaws due to our lack of knowledge about what an ASI could do, an ASI's ability to deceive us and inherent human bias in the risk process.

These issues are not unique to ASIs; they are also apparent when dealing with dynamic malicious threat actors such as professional criminals and terrorists. However, the potential for devastating harm to humanity is unique to ASIs.

The recognition that current risk management is an incomplete approach to ASIs needs to occur and in the absence of alternative methodologies the baseline assumption should be that powerful AI will be malicious and that maliciousness need to be managed using a data-driven approach of monitoring, guiding and managing intent.

The very real dangers of ASIs are not a reason to stop their development. Firstly, it is unlikely any global effort to stop ASI development would be successful and pushing ASI development underground would likely have a negative effect on overall safety. ASIs have enormous potential to help humanity and failing to develop safe ASI is in itself an existential threat called a technological arrest (Bostrom 2002).

[13] There are numerous statistical tools to detect abnormal deviation within a process. The Shewhart chart is for example purposes only as it is a well-known and well tested model. Much more advanced process control tools such as an exponentially weighted moving average could also be used. The statistical model used will entirely depend on factors such as the type of data being measured, the subgrouping of data, the volume of data and the type of data deviation it is trying to detect.

Secondly, there may not be a practical technological solution to stop the development of ASI. Constraint-based solutions such as restricting computing power through hardware constraints, software constraints and the introduction of various computing biases as a means to create "artificial stupidity" in AI may face insurmountable technical challenges, with Trazzi and Yampolskiy noting *"prohibiting the AGI from hardware or software self-improvement might prove a very difficult problem to solve and may even be incompatible with corrigibility"* (2018). This issue is apparent with the proposed treachery threat management model, if an AI behaves in an unexpected way there may be no safe way to bring it back online with any degree of certainty that it is not going to continue to develop capabilities and probe for vulnerabilities. The issue of safe restoration of misbehaving AI's needs further development.

Risk managers in any organisation currently developing AI need to be actively aware of the methodological failings of their approach and work with developers to ensure that the concept of guiding and blocking malicious intent through data-driven models is deeply embedded within the development roadmap.

Asimov's 3 Laws of Robotics made famous in his 1942 short story "Runaround" was an early acknowledgement of the need for safety protocols in ASIs (Asimov 1942). 75 years later we are closer to the reality of needing them, and yet we still harbour an irrational belief that as humans we can somehow outsmart an intelligence that could be millions of times greater than ours. Risk models for AI need to shift from static, anthropomorphic models and focus on data-driven models to measure intent, manage intent and prevent the treacherous turn.

Acknowledgements The author thanks both Nick Bostrom and Eliezer Yudkowsky for their work in garnering serious attention to the risks of superintelligence. Their works on the risks of artificial superintelligence are central to this paper. Also, thanks to Dr James Bradley and Professor Emeritus Denise Bradley AC for their editorial feedback and both Dr Paul Baldock, Ben Cornish and David Harding for their technical assistance.

References

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. https://arxiv.org/abs/1606.06565. Accessed 02 Mar 2019

Asimov I (1942) I, Robot. Doubleday, New York

AXELOS Limited (2014) Managing successful projects within PRINCE2®. AXELOS Limited, London

Bailey R (2017) Will superintelligent machines destroy humanity? http://reason.com/archives/2014/09/12/will-superintelligent-machines-destroy-h. Accessed 25 Sept 2017

Barrett JL (1996) Conceptualizing a nonnatural entity: anthropomorphism in god concepts. Cogn Psychol 31(3):219–247

Baum SD, Goertzel B, Goertzel T (2011) How long until human-level AI? Results from an expert assessment. Technol Forecast Soc Change 78(1):185–195

Baumann T (2017) Reducing risks of future suffering. http://s-risks.org/focus-areas-of-worst-case-ai-safety/. Accessed 9 Nov 2016

BBC (2014) Stephen Hawking warns artificial intelligence could end mankind. http://www.bbc.com/news/technology-30290540. Accessed 20 Aug 2017

Bostrom N (2002) Existential risks: analyzing human extinction scenarios and related hazards. J Evol Technol 9(1):1–37

Bostrom N (2006) How long before superintelligence? Linguist Philos Investig 5(1):11–30

Bostrom N (2014) Superintelligence. Oxford University Press, Oxford

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B et al (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. arXiv: Artificial Intelligence

Chalmers DJ (2010) The singularity: a philosophical analysis. J Conscious Stud 17:7–65

Chun Y, Lee J-L (2013) Traces of occupancy and its effect upon burglar's residential target selection. Soc Sci 2(3):135–141

Cisco (2019) AI Ops and the future of performance monitoring. https://blogs.cisco.com/cloud/our-vision-for-aiops-the-central-nervous-system-for-it. Accessed 11 Mar 2019

Cohen E (2017) The definitive guide to project management methodologies. https://www.workamajig.com/blog/project-management-methodologies. Accessed 02 Mar 2019

Danaher J (2014) Philosophical disquisitions. https://philosophicaldisquisitions.blogspot.com/2014/07/bostrom-on-superintelligence-3-doom-and.html. Accessed 27 Feb 2019

Ding F, Ge Q, Jiang D, Fu J, Hao M (2017) Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. PLoS ONE 12:e0179057

Dowd M (2017) Elon Musk's billion-dollar crusade to stop the A.I. apocalypse. https://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x. Accessed 18 Dec 2017

Facebook (2019) Facebook AI research. https://research.fb.com/category/facebook-ai-research/. Accessed 04 Mar 2019

Future of Life Institute (2017) Future if life institute. https://futureoflife.org/autonomous-weapons-open-letter-2017/. Accessed 5 Nov 2017

GE Power Systems University (2001) Six sigma black belt book of knowledge, 4th edn. GE Power Systems University, New York

Goertzel B (2014) Toward a formal model of cognitive synergy. OpenCog Foundation, Hong Kong

Goertzel B (2015) Superintelligence: fears, promises and potentials. J Evol Technol 24(2):55–87

Goertzel B, Pitt J (2012) Nine ways to bias open-source AGI toward friendliness. J Evol Teachnol 22(1):116–131

Google (2018) AI at Google: our principles. https://www.blog.google/technology/ai/ai-principles/. Accessed 04 Mar 2019

Grace K (2015) Less wrong. https://www.lesswrong.com/posts/B39GNTsN3HocW8KFo/superintelligence-11-the-treacherous-turn. Accessed 27 Feb 2019

Harding D (2014) Threat management: the coordinated focus on the threat actor, their intentions, and attack cycle. J Appl Secur Res 9(4):478–494

Harding D (2016a) Security management: a dangerously overrated and broken pardigm. http://www.criminalthreatmanagement.com/start-here/. Accessed 11 Apr 2017

Harding D (2016b). Security risk management: ineffective at best, dangerous at worst. http://www.securitysolutionsmagazine

.biz/2016/11/09/security-risk-management-ineffective-at-best-dangerous-at-worst/. Accessed 4 Nov 2017
Heemstra FF, Kusters RR, Man HD (2003) Guidelines for managing bias in project risk assessment. https://narcis.nl/publication/recordid/oai:library.tue.nl:671251. Accessed 5 Nov 2017
Hoorens V (1993) Self-enhancement and superiority biases in social comparison. Eur Rev Soc Psychol 4(1):113–139
Hubbard DW (2015) The failure of risk management: why it's broken and how to fix it. Wiley, New York
Hudsal J (2015) Risk management—vocabulary. http://www.husdal.com/2010/11/21/risk-management-vocabulary/. Accessed 7 Nov 2017
International Organisation for Standardisation (2018) International standard ISO 31000: risk management—guidelines. ISO 31000:2018 risk management-guidelines
Jamali G, Oveisi M (2016) A study on project management based on PMBOK and PRINCE2. Math Models Methods Appl Sci 10(6):142
Karaman E, Kurt M (2015) Comparison of project management methodologies: prince 2 versus PMBOK for it projects. Int J Appl Sci Eng Res 4(4):572–579
Kimball RC (2000) Failures in risk management. N Engl Econ Rev 15(Issue January):3–12
Kruger J, Dunning D (1999) Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 77(6):1121–1134
Li Z, Sun D, Li B, Li Z, Li A (2018) Terrorist group behavior prediction by wavelet transform-based pattern recognition. Discrete Dyn Nat Soc 2018:1–16
Lowder J (2010) Why the "risk = threats × vulnerabilities × impact" formula is mathematical nonsense. https://www.bloginfosec.com/2010/08/23/why-the-risk-threats-x-vulnerabilities-x-impact-formula-is-mathematical-nonsense/. Accessed 11 Nov 2017
Meek T, Barham H, Beltaif N, Kaadoor A, Akhter T (2016) Managing the ethical and risk implications of rapid advances in artificial intelligence: a literature review. http://ieeexplore.ieee.org/document/7806752. Accessed 5 Nov 2017
Moran A (2014) Agile risk management (2014 edition). Springer, Berlin
Office of Government Commerce (2007) The official introduction to the ITIL service lifecycle. TSO (The Stationery Office), UK
OpenAI (2019a) About OpenAI. https://openai.com/about/#mission. Accessed 04 Mar 2019
OpenAI (2019b) Gym. http://gym.openai.com/. Accessed 05 Mar 2019
Pallier G, Wilkinson R, Danthiir V, Kleitman S, Knezevic G, Stankov L et al (2002) The role of individual differences in the accuracy of confidence judgments. J Gen Psychol 129(3):257–299
Pande PS, Neuman RP, Cavanagh RR (2007) The six sigma way. https://link.springer.com/10.1007/978-3-8349-9320-5_24. Accessed 9 Nov 2017
Pennycook G, Ross RM, Koehler DJ, Fugelsang JA (2017) Dunning–Kruger effects in reasoning: theoretical implications of the failure to recognize incompetence. Psychon Bull Rev 24(6):1774–1784
PMBOK® Guide—Sixth Edition (2017) PMBOK® Guide—Sixth Edition. Project Management Institute, Newtown Square
Power M (2009) The risk management of nothing. Acc Organ Soc 34:849–855
RAND National Security Research Division (2005) "Connecting the dots" in intelligence detecting terrorist threats in the out-of-the-ordinary. https://www.rand.org/pubs/research_briefs/RB9079/index1.html. Accessed 27 Feb 2019
Raz T, Hillson D (2005) A comparative review of risk management standards. Risk Manag 7(4):53–66
Roper C (1999) Risk management for security professionals, 1st edn. Butterworth-Heinemann, Oxford
Sabato G (2009) Financial crisis: where did risk management fail? https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1460762. Accessed 5 Nov 2017
Salih T, Mohammad K, Zhuang J (2017) New framework that uses patterns and relations to understand terrorist behaviors. Expert Syst Appl 78:358–375
Schneider CM, Moreira AA, Andrade JS, Havlin S, Herrmann HJ (2011) Mitigation of malicious attacks on networks. Proc Natl Acad Sci 108:3838–3841
Smith L, Louis E (2010) Cash in transit armed robbery in Australia. http://aic.gov.au/media_library/publications/tandi_pdf/tandi397.pdf. Accessed 5 Nov 2017
Sydney Morning Herald (2017) Be extremely afraid: Elon Musk has a grim warning for US governors. http://www.smh.com.au/world/be-extremely-afraid-elon-musk-has-a-grim-warning-for-us-governors-20170716-gxc3yy.html. Accessed 25 Aug 2017
Taylor EZ, Blaskovich J (2011) By the numbers: individual bias and enterprise risk management. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2504164. Accessed 5 Nov 2017
Trazzi M, Yampolskiy RV (2018) Building safer AGI by introducing artificial stupidity. arXiv: Artificial Intelligence
Viljoen D, Musvoto SW (2013) The fluctuating nature of risk management models. http://dspace.nwu.ac.za/handle/10394/11332. Accessed 5 Nov 2017
Wang JX, Roush ML (2000) What every engineer should know about risk engineering and management. Marcel Dekker, New York
Wood M (2002) Statistical process monitoring in the 21st century. https://researchgate.net/profile/michael_wood15/publication/237227899_statistical_process_monitoring_in_the_21st_century/links/02e7e539971cddef73000000.pdf?inviewer=true&pdfjsdownload=true&disablecoverpage=true&origin=publication_detail. Accessed 20 Mar 2019
Yampolskiy R (2012) Leakproofing the singularity artificial intelligence confinement problem. J Conscious Stud 19:194–214
Yudkowsky E (2008) Artificial intelligence as a positive and negative factor in global risk. https://intelligence.org/files/AIPosNegFactor.pdf. Accessed 23 Oct 2017
Yudkowsky E (2015) Rationality: from AI to zombies. Machine Intelligence Research Institute, Berkeley