The AAAI-17 Workshop on
AI, Ethics, and Society
WS-17-02

The Formalization of AI Risk Management and Safety Standards

Shabnam Ozlati
Human Factors Consulting Services, Inc. *
shabnam@hfcsi.com

Roman Yampolskiy
Computer Engineering and Computer Science, University of Louisville
roman.yampolskiy@louisville.edu

Abstract

Researchers have identified a number of possible risks posed to humanity by anticipated advancements in artificial intelligence (AI), but the extant literature on the topic is largely academic or theoretical in nature. Despite the likelihood that much of AI's future development will occur in industry settings, the insights generated by the AI safety research community have yet to be translated into a set of practical guidelines for working developers, project managers, and other industrial stakeholders. There are no currently established standards in place to guide the safe development of AI technologies, but the risk management approach employed in mature industries such as aerospace and medical manufacturing offers a promising model that may be adapted to AI related safety concerns. Within these industries, the safety guidelines and best practices derived from the risk management approach are developed, evaluated, formalized, and disseminated by industry specific Standards Developing Organizations (SDOs). This paper proposes a project to spur the development and adoption of formal AI risk management practices by demonstrating the approach's viability through the completion of an AI risk assessment process. The results of the proposed activities are intended to lay the initial groundwork necessary for the eventual creation of an AI SDO.

1 Introduction

This paper outlines a proposed plan aimed towards adapting longstanding industry practices (e.g., aerospace and medical) to long-term AI safety concerns in order to spur the development of a formal risk management approach focusing on the risk assessment step of the risk management process (risk identification, analysis, and evaluation). The results of the planned risk assessment activities could then be used as a starting point for efforts towards the creation of a standing body that can develop, evaluate and disseminate AI safety standards.

2 General Background & Rationale

Much of AI's future development will likely occur in industry, yet the majority of extant work on long-term AI Safety concerns has an academic or theoretical focus. This implies that as the technology advances, developers may create increasingly powerful and potentially dangerous AI systems without recourse to a generally accepted set of safety best practices, standards and process guidelines for the verification and validation of new systems. Therefore, it is critical to establish concrete means by which expert judgment can be translated into practical guidance for system designers and other institutional stakeholders who will play a role in the development of future AI systems.

As a model for how to effectively develop and disseminate practical safety guidance for AI developers, it is instructive to look at hi-tech industries such as aerospace and medical devices that must verify and validate systems that can cause serious adverse consequences as a result of failure or misuse. It is standard practice in such industries to 1) employ formal risk management processes, and 2) develop and disseminate consensus standards through accredited Standards Developing Organizations.

3 Methodology Rationale

3.1 Focus on Artificial General Intelligence (AGI)

There are a number of experts currently working on identifying the risks posed by the development of advanced AI, and generating proposed safeguards against those risks. The inclusivity of the term AGI allows for an accounting of potential risks posed by a wide variety of possible future AI systems, and provides a framework that is broad enough to accommodate this diversity.

Due to an AGI's generalized capabilities and potentially diffuse set of goals, the potential for unpredictable, self-directed behavior may create more risks to humanity than narrow AI. In order to maintain a manageable scope while providing the maximum long-term benefit to human safety and wellbeing, the proposed research will not address issues related to narrow AI.

3.2 Modified Delphi Method

The risk management approach currently employed in a number of industries involves the formation of a consensus among a given organization's experts and stakeholders. This process by which this consensus is reached utilizes a framework based on broader, industry wide standards that are themselves the product of a national and international consensus. The absence of an expert consensus regarding both the risks posed by AGI and the proper measures to manage those risks impedes the critical task of creating a set of practical guidelines for the development of the technology. Because it is designed to elicit an expert consensus while reducing bias and noise, the Delphi methodology has been selected for the creation of a risk assessment table. The risk assessment table's format has been adapted from similar documents that are used in high tech, high risk industries to organize the collective expert judgment generated by a thorough risk management process.

The Delphi method was first developed by the RAND Corporation in the 1950's as a way of arriving at an expert consensus while minimizing the problems of group think, entrenched opinion, and politically motivated manipulation (Rowe & Wright, 1999). The method works by harnessing the distributed expertise of multiple experts under conditions of respondent anonymity. In addition to respondent anonymity, the impact of interpersonal and professional political considerations on judgment is attenuated by the fact that the panelists do not communicate directly with one another. This method has been shown to produce accurate predictions relative to other forms of forecasting (Rowe & Wright, 1999), and can facilitate theory building and conceptualization as well (Okoli & Pawlowski, 2004).

Traditionally, the first round of a Delphi study involves collecting expert responses to an open ended inquiry on the subject under study. These responses are then aggregated and utilized in the design of the structured questionnaire employed in the second round of data collection. In subsequent rounds, panelists are provided with a "group" answer aggregated from the previous round's responses, and asked to provide feedback and corrections. This iterative feedback process continues until the group answer reaches an acceptable degree of stability.

3.3 Recruitment, Selection, and Retention of Panelists

Panelists could be recruited based on demonstrated expertise in the field of AI, computer science, or other relevant disciplines. The drafting of the list of potential candidates will be informed by publically available information, as well as consultation with members of organizations such as MIRI, FHI, FLI, and others.

Brian Tomasik (2014) has noted systematic differences in the professional backgrounds of the well-known public proponents of hard and soft takeoff scenarios for the development of advanced AI. Panelists should be recruited from both theoretical and applied backgrounds who represent the spectrum of opinion on the matter. A diverse panel of experts should serve to expand on the range of inputs available for the generation of safeguards, as thinkers from outside of a given school of thought may take novel approaches to a given problem.

Currently, the takeoff speed of future AI systems is unknown, so it may be helpful to encourage best practices that can account for a range of possible outcomes (Sotala & Yampolskiy, 2013). Some suggest that safeguards against the hazards presented by a soft takeoff may provide additional time to develop anticipatory responses to a hard takeoff, even if their direct efficacy against hard takeoff related hazards is in question.

3.4 Conducting Separate Risk Assessments for Different System Types

It is likely that developers will employ a variety of approaches in order to make progress towards more powerful AI. Different categories of AI system may present distinct risks, and may require mitigation strategies that are not applicable to systems with differing characteristics. To generate a systematic risk assessment for future AI systems and demonstrate the general viability of the risk management approach, and produce a central repository for possible AI related risks, the proposed study must account for the diversity of AI design.

Therefore, the Delphi study should entail separate risk assessments for different system categories. For the purposes of this study, systems may be classified based on a number of factors relevant to their potential behavior, including intended use (such as an oracle system), architecture (human brain emulation, distributed artificial intelligence etc.), or programmed goals (such as a value learning machine). An initial list of possible system types will be drafted prior to the first round of the Delphi study, but expert panelists will be instructed to add a category if they find the list incomplete.

3.5 Conducting Risk Assessments for Hard and Soft Takeoff Scenarios

Among experts, there is little consensus as to the nature and probability of the risks associated with AGI, and disagreement as to how these risks might be mitigated. One of the factors that drives this lack of agreement is the lack of a shared set of assumptions about the probable trajectory of AGI's development. There is considerable controversy concerning even the possibility of the creation of AGI, and among those who do predict its eventual existence, there is little agreement as to the way in which its existence will come to pass. Proponents of the soft takeoff propose that AGI will develop slowly through incremental improvements in software and hardware. According to this predicted trajectory, development may occur at a gradual enough pace for corrective measures to be taken once advanced AI has been achieved. The hard takeoff scenario, on the other hand, involves a burst of rapid AGI development that will quickly outpace attempts to control the direction of the technology's growth and behavior. For an in depth discussion of takeoff speed and risk, see Sotala and Yampolskiy (2013).

The Delphi method's efficacy is based on an iterative approach that allows experts to revise their initial positions in light of the feedback provided by other experts. Given the potentially irreconcilable sets of assumptions that undergird predictions of the hard/soft takeoff of AI, it may be unrealistic to expect adherents of either camp to jettison their assumptions, even in the face of feedback. To prevent this from becoming a bottleneck, experts could be asked to make two separate sets of AGI risk assessments based on two hypothetical futures: One in which soft-take off occurs, and one in which a hard takeoff occurs. Because the exercise would allow experts to generate one set of assessments based on their own set of assumptions about takeoff speed, it would allow them to temporarily adopt the opposing framework without relinquishing their own claims.

This dual scenario approach may also bring fresh insights to light that might not otherwise discovered. Because this exercise could encourage panelists to temporarily adopt a new intellectual framework, it would allow them to consider issues that they may not have previously grappled with and make new contributions (Ramirez & Wilkinson, 2016). For example, a convinced proponent of the soft takeoff scenario may be capable of generating novel solutions to the risks posed by a hard takeoff, but may have previously lacked the motivation to do when laboring under a position that holds such a scenario unlikely.

Additionally, there may be added value in encouraging thinkers who predict a soft takeoff to give thought to the hazards presented by a hard takeoff. Research has shown that the ability to generate ideas may be enhanced when thinking about psychologically distant events or objects (Forster, Friedman, & Liberman, 2004; Jia, Hirt, & Karpen, 2009). Probability has been shown to be a form of psychological distance (Trope & Liberman, 2010). This psychological phenomenon may imply that panelists' thinking will be enhanced when considering responses to outcomes to which they assign a lower probability.

3.6 Avoiding Timeline Based Predictions

Muller and Bostrom's recent survey of AI experts (2014) illustrates very well the disparate nature of expert predictions of the timeline of AI development. When asked when machines will be capable of carrying out most human professionals at least as well as a typical human can, survey respondents gave answers ranging from a decade from the time of the survey, to never. Other researchers (Armstrong, Sotala, & ÓhÉigeartaigh, 2014) have shown that experts have historically generated inaccurate predictions of the timeline of AI development. In order to remove this source of contention and inaccuracy, the study proposed here would avoid issues of predicting the timeframe of AI's development. A secondary rationale for this decision is the fact that the AI's long term impact on humanity may not dependent on the exact timing of its creation.

4 Definition of Terms

In order to fit the proposed AI safety research activities within the broader framework of the formalized risk management practices and safety standards employed in other industries, currently used risk management nomenclature could be applied to AI safety. Consequently, the terminology used will adhere to the definitions provided in ANSI/AAMI/ISO 14971:2007, which are based on internationally recognized definitions found in documents such as ISO 9000:2005 and ISO/IEC Guide 51:1999. For the sake of readers unfamiliar with these documents, definitions have been provided below.

Harm: injury or damage to the health of people, or damage to property or the environment

Hazard: Potential source of harm

Hazardous Situation: Circumstance in which people, property, or the environment are exposed to one or more hazards.

Severity: measure of the possible consequences of a hazard

Risk: Combination of the probability of occurrence of harm and the severity of that harm

Risk Analysis: Systematic use of available information to identify hazards and estimate the risk

Risk Estimation: Process used to assign values to the probability of occurrence of harm and the severity of that harm

Risk Assessment: Overall process comprising a risk analysis and a risk evaluation
Risk Evaluation: process of comparing the estimated risk against given risk criteria to determine the acceptability of the risk.
Residual Risk: Risk remaining after risk control measures have been taken
Risk management: systematic application of management policies, procedures, and practices to the tasks of analyzing, evaluating, controlling, and monitoring risk
Risk control: Process in which decisions are made and measures implemented by which risks are reduced to, or maintained within, specified levels
Inherently safe design: measures taken to eliminate hazards and/or to reduce risks by changing the design or operating characteristics of the product or system

5 AI Safety Standards Developing Organization (AISSDO)

The activities sought in this paper are intended to satisfy a number of necessary preconditions for the creation of a recognized set of AI safety standards to be drafted under the auspices of a fully staffed, accredited AI safety Standards Developing Organization (SDO). An SDO is a non-governmental organization that is accredited by the American National Standards Institute (ANSI) to develop and maintain national standards. National standards, in turn, can contribute to the formation of international standards through other accredited bodies (American National Standards Institute, 2010). Additionally, the formulation of safety standards requires a full risk management process that entails risk assessment, risk control and risk monitoring. The results of the proposed risk assessments can be used to inform the future drafting of said standards, but do not represent the only necessary input for the process. While the activities described in this paper are intended to generate value in the present, the overarching goal of the current project is to provide the initial impetus for the longer term, more ambitious project of institution building. It is worth mentioning that currently several processes are ongoing such as the IEEE Initiative on Autonomous Systems to address the need for building an AISSDO (IEEE Standards Association, 2016). The critical task of guiding industry towards safe practices will be an ongoing endeavor that cannot be completed without further coordinated action by concerned parties, but it is feasible to encourage developments in that direction.

6 Conclusion

If successful, the integration of AI safety into the broader risk management framework could benefit industry and society as a whole. The systematic, consensus driven identification, rating, and classification of hazards may reduce the likelihood that an unacceptable risk will be overlooked. It will also foster innovation in AI by providing large, risk-averse organizations with an institutionally usable template for formal safety efforts, thereby facilitating the governance of AI projects and allowing considerable resources to flow into AI research.

While there is value in generating consensus on the hazards generated by future AI and the proper responses to mitigate them, the uncertainty inherent in predicting the nature and timing of future developments must be acknowledged. It is clear that any consensus developed utilizing even the most complete current understanding of the topic will need to be revised in the future in light of information that is not yet available. Therefore, the proposed research aims to provide benefits that are not contingent upon the exact nature or timing of currently unknowable future events. The risk management process that we aim to undertake can serve as a usable template for other AI researchers and developers, even if the consensus it generates is subject to change. The dissemination of a standardized, systematic method for addressing the potential risks posed by AI systems is likely to encourage developers to dedicate more resources to AI safety. Because it offers the possibility of increasing the total amount of labor directed towards the mitigation of AI related risk, the creation of a usable, concrete risk management template may help to address the nearsightedness problem (Ord, 2014) that can affect efforts to prevent future existential risk.

The creation of a usable risk management framework for AI may be of long term benefit to industry, the research community, and humanity as a whole, but the proposed research is also intended to be useful in the immediate future. Succinct, accessible documents that summarize and delineate the risks posed by future AI systems and potentially viable mitigations could serve as a useful introduction to the existentially critical field of AI safety for the lay community. In the age of increasingly powerful computing, the human race is confronted with challenges and opportunities that are currently outside of the awareness of vast swaths of the public.

Policymakers, philanthropists, educators, business leaders, aspiring specialists, and even concerned members of the public could play a role in shaping our species' response to the challenges of safe AI, yet they may lack the awareness required for the task. It may be impractical to expect non-specialists to wade through the literature to acquire a basic grasp of the relevant issues. It is our hope that the freely available publications generated by the proposed study will serve as powerful educational tools for the outreach efforts undertaken by others in the AI safety community.

7 References

American National Standards Institute. 2010. Overview of the U.S. Standardization System: Voluntary Consensus Standards and Conformity Assessment Activities - 3rd ed.

Armstrong, S., Sotala, K., & ÓhÉigeartaigh, S. S. 2014. The errors, insights and lessons of famous AI predictions and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence 26(3): 317-342.

Association for the Advancement of Medical Instrumentation. ANSI/AAMI/ISO 14971: 2007. Medical devices - Application of risk management to medical devices.

Forster, J., Friedman, R.S., & Liberman, N. 2004. Temporal construal effects on abstract and concrete thinking: consequences for insight and creative cognition. Journal of Personality and Social Psychology 87:177-189.

Jia,L., Hirt, E. R., & Karpen, S. C. 2009. Lessons from a Faraway land: The effect of spatial distance on creative cognition. Journal of Experimental Social Psychology 45:1127-1131.

Müller, V. C., & Bostrom, N. 2014. Future progress in artificial intelligence: A survey of expert opinion. Fundamental Issues of Artificial Intelligence.

Okoli, C., & Pawlowski, S. D. 2004. The Delphi method as a research tool: an example, design considerations and applications. Information & Management 42(1): 15-29.

Ord, T. 2014. The timing of labour aimed at reducing existential risk. http://www.fhi.ox.ac.uk/the-timing-of-labour-aimed-at-reducing-existential-risk/

Ramírez, R., & Wilkinson, A. 2016. Strategic Reframing: The Oxford Scenario Planning Approach.: Oxford University Press.

Rowe, G., & Wright, G. 1999. The Delphi technique as a forecasting tool: issues and analysis. International journal of forecasting 15(4): 353-375.

Sotala, K., & Yampolskiy, R. V. 2013. Responses to catastrophic AGI risk: A survey (Technical Reports, 2013 (2)). Berkeley, CA: Machine Intelligence Research Institute.

IEEE Standards Association. 2016. The IEEE Global Initiative for Ethical and Autonomous Systems. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Tomasik, B. 2014. Predictions of AGI Takeoff Speed vs. Years Worked in Commercial Software.

http://reducing-suffering.org/predictions-agi-takeoff-speed-vs-years-worked-commercial-software/

Trope, Y., & Liberman, N. 2010. Construal-level theory of psychological distance. Psychological review 117(2): 440.