Risk Analysis and Risk Management for the Artificial Superintelligence Research and Development Process

Anthony M. Barrett and Seth D. Baum
Global Catastrophic Risk Institute
http://sethbaum.com * http://tony-barrett.com * http://gcrinstitute.org

Abstract

Artificial superintelligence (ASI) is increasingly recognized as a significant future risk. In the absence of adequate safety mechanisms, an ASI may even be likely to cause human extinction. Thus ASI risk scenarios merit attention even if their probabilities are low. ASI risk can be addressed in at least two ways: by building safety mechanisms into the ASI itself, as in ASI safety research, and by managing the human process of developing ASI, in order to promote safety practices in ASI research and development (R&D). While ASI researchers and developers typically do not intend to cause harm through their work, harm may nonetheless occur due to accidents and unintended consequences. Thus opportunities may exist to reduce ASI risk through engagement with the R&D process. This paper surveys established methodologies for risk analysis and risk management, emphasizing fault trees and event trees, and describes how these techniques can be applied to risk from ASI R&D. A variety of risk methodologies have been developed for other risks, including other emerging technology risks, but their application to ASI has thus far been limited. Insights from risk literatures could improve on what existing analyses of ASI risk have yet been conducted. Likewise, a more thorough and rigorous analysis of ASI R&D processes can inform decision making to reduce the risk. The decision makers include governments and non-governmental organizations active in ASI oversight, as well anyone conducting ASI R&D. All of these individuals and groups have roles to play in addressing ASI risk.

1. Introduction

A substantial amount of work has made the case that global catastrophic risks (GCRs) deserve special attention (Sagan 1983; Ng 1991; Bostrom 2002; Beckstead 2013; Maher Jr. and Baum 2013). Major issues in addressing GCRs include assessing the probabilities of such catastrophic events and assessing the effectiveness and tradeoffs of potential risk-reduction measures in light of limited risk-reduction resources and tradeoffs in using them.

Certain types of artificial intelligence (AI) have been proposed as a potentially large factor in GCR. One specific AI type of great concern is artificial superintelligence (ASI), in which the AI has intelligence vastly exceeding humanity's across a broad range of domains (Bostrom 2014). ASI could potentially either solve a great many of society's problems or cause catastrophes such as human extinction, depending on how the ASI is designed (Yudkowsky 2008) .

The AIs that exist at the time of this writing are not superintelligent, but ASI could be developed sometime in the future. It is important to consider the long-term possibilities for ASI in order to help avoid ASI catastrophe. With careful analysis, it may be possible to identify indicators that ASI development is going in a dangerous direction, and likewise to identify risk

management actions that can make ASI development safer. However, long-term technological forecasting is difficult (Lempert et al. 2003), making ASI risks difficult to characterize and manage. Additional challenges come from the possibility of ASI development going unnoticed (such as in covert development projects) and from weighing the risks posed by ASI against the potential benefits that ASI could bring.

This paper surveys established methodologies for risk analysis and risk management as they can be applied to ASI risk. ASI risk can be addressed in at least two ways: (1) by building safety mechanisms into the ASI itself, as in ASI "Friendliness" research, and (2) by managing the human process of researching and developing ASI, in order to promote safety practices in ASI research and development (R&D). This paper focuses on the human R&D process because it has similarities to the R&D processes for other emerging technologies. Indeed, the ASI risk analysis ideas presented here are similar to our own work on risks posed by another emerging technology, synthetic biology (Barrett 2014).

The ultimate goal of ASI risk analysis is to help people make better decisions about how to manage ASI risks. Formalized risk methodologies can help people consider more and better information and reduce cognitive biases in their decision making. A deep risk perception literature indicates that people often have grossly inaccurate perceptions of risks (Slovic et al. 1979). One example is in perceptions of "near miss" disasters that are luckily but narrowly avoided. An individual's framing of the near miss as either a "disaster that did not occur" or a "disaster that almost happened" tends to decrease or increase, respectively, their perception of the future risk of such a disaster (Dillon et al. 2014). This, combined with the high stakes of ASI, suggests substantial value in formal ASI risk analysis.

2. Key ASI R&D Risk and Decision Issues

For risk analysis, important questions concern the probabilities, timings, and consequences of the invention of key ASI technologies. Regarding the consequences, Yudkowsky (2008), Chalmers (2010) and others argue that ASIs could be so powerful that they will essentially be able to do whatever they choose. Yudkowsky (2008) and others thus argue that technologies for safe ASI are needed before ASI is invented; otherwise, ASI will pursue courses of action that will (perhaps inadvertently) be quite dangerous to humanity. For example, Omohundro (2008) argues that a superintelligent machine with an objective of winning a chess game could end up essentially exterminating humanity because the machine would pursue its objective of not losing its chess game, and would be able to continually acquire humanity's resources in the process of pursuing its objective, regardless of costs to humanity. We refer to this type of scenario as an ASI catastrophe and focus specifically on this for the remainder of the paper.

The risk of ASI catastrophe has the dynamics of a race. Society must develop ASI safety measures before it develops ASI, or else there will be an ASI catastrophe. Estimating ASI catastrophe risk thus requires estimating the probabilities of ASI and ASI safety measures occurring at different times. For ASI invention, a number of technology projection models exist, e.g. The Uncertain Future (Rayhawk et al. 2009a). ASI safety measure models are less well formulated at this point but would be needed for a complete risk analysis.

For risk management, the most important question is: What policies (public or private) should be pursued? A variety of ASI risk reductions policy options have been identified (e.g., Sotala and Yampolskiy 2015; a version of which appears as Part 1 of this volume). At least three sets of policies could be followed, each with its own advantages and disadvantages:

    1) Governments, corporations, and other entities could implement ASI R&D regulations

within their jurisdictions, and pursue treaties or trade agreements for external cooperation. Regulations could restrict risky ASI R&D. However, implementation could be costly and could impede benign R&D. It would also be unlikely to be universally agreed and enforced, such that risky research could proceed in unregulated regions or institutions.

2) Security agencies could covertly target risky ASI projects. Similar covert actions have reportedly been taken against other R&D projects, such as the Stuxnet virus used against Iran's nuclear sector. Such actions can slow down dangerous projects, at least for a while, but they could also spark popular backlash, harden project leaders' desire to continue, and provide dangerous ASI R&D efforts with incentives to avoid detection.

3) Governments, corporations, foundations, and other entities could fund ASI safety measure development. This could increase the probability of ASI safety measures being available before ASI. However, ASI communities do not have consensus on ASI safety measure concepts or best approaches – more on this below – and some ASI safety measures may still take more time to develop than ASI, in which case ASI catastrophe would still occur.

Here are some potentially important factors:

- Provability of safety measures built into ASI goals ("Friendliness") (Muehlhauser 2013)
- ASI technology development arms race (Shulman and Armstrong 2009)
- Whether and how a government ought to support ASI safety (McGinnis 2010)
- Government intervention blowback risks and other drawbacks (Chalmers 2010 footnote 14)
- Potential for "hard takeoff" vs. "soft takeoff" of AI, and their relation to hardware vs. software limitations on AI takeoff (Shulman and Sandberg 2010)

3. Risk Analysis Methods

3.1 Fault Trees

Fault trees represent the ways that events and conditions could combine to lead to a particular outcome. Each node in the tree represents a particular event, such as an attempted use of ASI, or a condition, such as the existence of a new ASI technology. The "top event" node in the tree represents the scenario outcome. Below the top node, the tree branches out with additional nodes. Each layer in the tree represents the combination of events and conditions that could lead to the outcome in the layer directly above it. Nodes are connected by Boolean logic gates, such as OR, AND, and NOT gates, which are an important part of specifying the particular combinations of events and conditions that could result in the outcomes above them in the tree. The tree thus shows a set of possible scenarios, each of whose "fault" it could be for the occurrence of the top event.

Figure 1 presents the logic model for a simple ASI catastrophe scenario fault tree. ASI catastrophe occurs if building an ASI is physically possible, if an ASI is built, if the ASI gains "decisive strategic advantage", and if the ASI actions are unsafe. ASI can be built either (1) directly, meaning humans create ASI on their own, or (2) via recursive self-improvement, in which humans build a "seed" AI that builds successively more intelligent AIs until it becomes superintelligent. Decisive strategic advantage means "a level of technological and other advantages sufficient to enable it [the AI] to achieve complete world domination" (Bostrom 2014, p. 78). ASI actions are unsafe if the ASI uses its decisive strategic advantage to cause catastrophe. ASI actions are unsafe if (1) its goals are unsafe, such that it would cause catastrophe if it pursues its goals, and (2) it is not deterred by any humans, other AIs, or anything else, such that it pursues its goals.
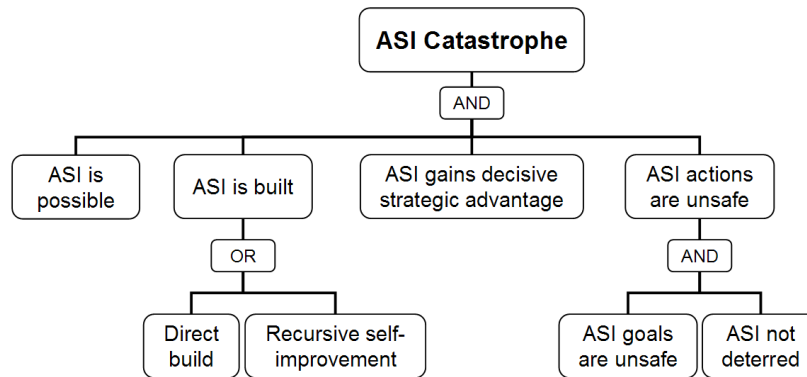
Figure 1: Simple ASI Catastrophe Fault Tree Logic Model

Fault tree logic models such as Figure 1 can be extended for quantitative risk analysis using parameters with any real-number value instead of just Boolean logic. Such fault trees can be used to estimate the occurrence rate or probability of the top-event outcome using other rate or probability variables as model inputs. Essentially, quantitative models are created by adding quantitative values for model parameters (often either rate or probability variables) in the fault tree logic models. Each node represents a variable with an associated rate (e.g. a rate of origination of entities that would attempt to create ASI if they obtain sufficient resources) or probability (e.g. a probability that a new ASI technology would be available to entities at a particular point in time). For mathematical details and an example of how rate and probability variables can be combined in a risk model fault tree, see Barrett et al. (2013). Quantitative models provide better comparison of risk magnitudes than logic models, but building quantitative models is harder, as it requires more data and/or assumptions regarding parameter values.

3.2 Event Trees

Event trees represent the possible outcomes of an event, and the probabilities of arriving at each of those outcomes. Event trees can be used to represent at least two kinds of important ASI catastrophe risk factors. The first and simplest application is to represent the probabilities of either the current (unknown) state, or the future state, of a particular condition. For example, an important ASI catastrophe risk factor is whether it is fundamentally possible to invent an ASI. Although some individuals may have opinions about whether the proposition is true, it is unknown at this point in time whether it is true. As shown in Figure 2, the probability of the proposition being true can be represented by the parameter p_T, and the probability of the proposition being false can be represented by the quantity $1 - p_T$.
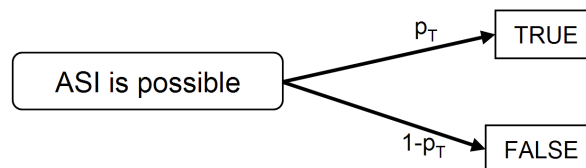

Figure 2: Event Tree of Whether ASI is Fundamentally Possible

A second application of event trees is for technology development modeling, working forwards from the current state of the world. ASI technology development models can provide estimates

of the probabilities of ASI development conditions as a function of time and potentially other variables (e.g. the financial resources of actors pursuing ASI).

Important conditions include (1) whether a technology has been invented or made available and (2) how affordable the technology is. These two conditions are for the forecasting of ASI developed via "grind" or "insight" (Armstrong and Sotala 2012). Grind involves applying established techniques repeatedly for gradual progress. Examples of grind include gradual progress in semiconductor manufacturing permitting faster hardware per dollar (as in Moore's law) and gradual progress in brain imaging permitting more detailed artificial brain emulations (as in whole brain emulation; see Sandberg and Bostrom 2008). Insight involves intellectual breakthroughs bringing fast, transformative progress. An example of insight is if ASI requires advances in algorithms that would not need advanced hardware to run on. Figure 3 shows an event tree for when an ASI "insight" breakthrough could occur, with each year between 2015 and 2100 as a possibility, and with a probability parameter for each year.
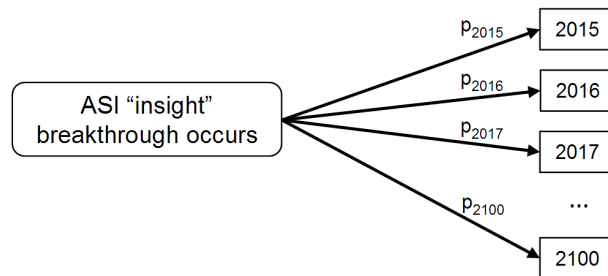

Figure 3: Event Tree of Year When "Insight" Breakthrough Occurs

3.3 Estimating Parameters for Fault Trees and Event Trees

Using fault trees and event trees for quantitative risk analysis requires estimating parameter values for all model components. Parameter estimation is straightforward when parameter values are known. However, parameter values are often uncertain, in which case techniques are needed for characterizing the uncertainty, typically in terms of probability distributions.

Many risk analyses form parameter probability distributions based on some combination of empirical data and expert judgment. Depending on the type of variable that the parameter represents, the parameter can often be approximated by one of several well-known probability distributions. For example, parameters representing rates or frequencies (holding any positive value) sometimes assume a Poisson process (e.g. for randomly-timed ASI creation events), in which case gamma distributions can be used; for parameters representing binomial processes (which have two possible outcomes: true/false, yes/no, etc.), beta distributions can be used (Bolstad 2007). For parameters representing multinomial processes (with more than two possible outcomes), Dirichlet distributions can be used.

For technological forecasting, linear regressions are often used to extrapolate trends, especially over the relatively short term, e.g. a period of one year (Millett and Honton 1991 pp. 13-14). Note that linear regressions can be used for transformations of raw data. The most notable example of this is the exponential progression seen in Moore's law: a plot of log(performance) vs. cost gives a fairly straight line (Stokes 2008). Lognormal distributions have been used for estimating invention date. For example, Rayhawk et al. (2009b) postulate a lognormal distribution for the number of years until there is enough brain imaging technology to build neuromorphic artificial general intelligence. Fallenstein (2013) suggests a Pareto distribution for ASI arrival date. Modis (2012) and Kurzweil (2012) debate the use of linear

regressions and logistic curves. Sandberg (2010) surveys a number of distributions and models that have been suggested in various ASI technology development models.

The probability distributions, whether based on empirical data or expert judgment, can be interpreted in Bayesian terms. Expert judgment can serve as a prior distribution, to be updated as empirical data is observed. In the absence of any expert judgment, uniform distributions can be used to represent ignorance about a parameter value, though care must be taken here because uniform distributions are sensitive to the choice of model structure (Pratt et al. 1995 pp. 236-237).

Finally, the passage of time can lead to further updating. In some cases, if time has passed and no new indicators have been observed, that can be counted as evidence for or against particular hypotheses, which itself should count as evidence for use in Bayesian updating.

3.4 Elicitation of Expert Judgment

Because ASI is an unprecedented technology, and because it may currently be at an early stage of development, there will be no empirical data for large portions of the parameters of any fault tree or event tree model of ASI R&D. To estimate these parameters with something more than just a uniform distribution, expert judgment is needed.

Expert judgment has significant limitations across all domains of expertise, including for AI predictions (Armstrong and Sotala 2012; Armstrong et al. 2014). However, expert elicitation best practices can help overcome the limitations. For example, actual experts should exist for the questions asked, and models or elicitation questions should be structured to take advantage of experts' knowledge (Morgan and Henrion 1990 pp. 128-137; Meyer and Booker 1991 pp. 24-26). Unfortunately, many expert judgments about AI have not used the best practices (Armstrong and Sotala 2012 sec. 4.1; Armstrong et al. 2014). For example, many AI predictions consist only of point estimates of when specific AI milestones will be accomplished; a more complete characterization of the uncertainty about AI milestone timing suggests using probability distributions instead of point estimates (Baum et al. 2011).

As argued by Armstrong and Sotala (2012), AI technological progress forecasts appear to have often been substantially wrong, even when made by "experts". However, ASI experts may not be the most knowledgeable individuals about technological forecasting, and vice versa. Thus better models for ASI technological forecasting may be constructed using expert elicitations of combinations of threat-domain experts and tech-forecasting experts as follows:

- Use ASI experts to inform model structure (e.g. the nature and number of major technological development steps necessary for ASI)
- Use technological forecasting experts to inform model parameters (e.g. quantities of time and resources typically required for major technological developments)

In addition, where possible, forecasting models should be structured to include intermediate-step claims for empirical testing and updating. That should help prevent excessive reliance upon experts without opportunities to check their forecasts.

3.5 Aggregation of Data Sources

In risk modeling, mathematical methods are often used to aggregate empirical data and expert judgment in order to arrive at a model that represents the best information available. However, such methods should not be used unthinkingly—there are multiple ways of conducting mathematical aggregation, and sometimes it is better to leave data disaggregated (Keith 1996). For example, differences in experts' views may result from important differences in their

fundamental assumptions; aggregating their views hides these differences, to the detriment of risk analysis and management. Morgan and Henrion (1990) advise using model analysis methods to identify the most important input factors in a model, and then to explore and communicate the key points about the model's dependence on values of key input factors, rather than simply trying to merge all input values together using aggregation methods.

In cases where aggregation of judgments from an expert elicitation process is appropriate, e.g. in combining differing opinions among experts with roughly similar fundamental assumptions, weighted aggregation is often performed using Bayesian statistics. We focus on weighting for beta distributions in the following. Basic Bayesian aggregation for beta distributions (for a binomial process) is described in Meyer and Booker (1991 pp. 331-335). The basic method assumes that each of several experts provides a judgment about beta distribution parameter values y and n, where y is the number of successes in n trials. Thus each expert i provides values y_i and n_i. Then aggregation of the judgments provides aggregated values y' and n', where $y' = \sum_i y_i$ and $n' = \sum_i n_i$. The basic approach also could be fairly simple to use for adjustable weighting of information from various sources. Weighting is used sometimes in processing expert elicitation judgments when not all experts are regarded as being equally credible. When one expert is viewed as being more credible than another, their views are given higher weighting. It is simple to extend the above-mentioned aggregation process to account for weights, by giving each expert's judgment a weight w_i. Then weighted aggregation of the judgments uses the formulas $y' = \sum_i w_i y_i$ and $n' = \sum_i w_i n_i$.

Expert judgment performance-based methods are sometimes used in deciding how much weight to give to judgments made by different experts in an expert elicitation, as in the expert judgment performance calibration methods of Cooke (1991) where the elicitor assesses the performance of experts based on how well they respond to questions with answers known to the elicitor (O'Hagan et al. 2006 pp. 184-185).

4. Risk Management Decision Analysis Methods

The point of risk analysis is, in general, not the analysis itself, but its potential to inform risk management decisions. For ASI catastrophe, there are many risk management options, which are available for a variety of decision makers (Sotala and Yampolskiy 2015). Relevant decision makers for ASI R&D include governments and non-governmental organizations active in ASI oversight, as well anyone sponsoring, conducting, or otherwise supporting ASI R&D. All of these individuals and groups have roles to play in addressing ASI risk. Ideally an ASI decision analysis would inform all of these many decisions, though in practice analysts must focus on only some decisions.

In simplest terms, risk management decisions are evaluated according to two factors: the options and the objectives. The options constitute the set of all possible risk management actions, including the act of doing nothing. In practice, there is often an infinity of possible actions. To make analysis tractable, one must select a finite portion of these options. When each option is evaluated by hand, a small number of options must be chosen, with each ideally representing some important class of options. For ASI R&D, important classes of options include abstaining from developing ASI ("relinquishment"; Joy 2000) and conducting ASI safety measures research.

The objectives for risk management decisions are the underlying goals or purposes that the decision makers seek to accomplish. The objectives can be expressed in terms of an objective function, optimization criterion, utility function, social welfare function, ethical framework, or

similar analytic paradigms. These paradigms have implicit commensurability between all items being valued (e.g. lives saved vs. dollars spent), which allows for a relatively simple equation for the expected value of a variety of activities types and their consequences (Clemen and Reilly 2001 p. 512). Other decision analysis research uses multi-criterion objective functions, seeking to identify options that perform well across a range of objectives (Keeney and Raiffa 1976).

ASI decision analysis is complicated by the prospect of including the ASI itself as an intrinsically valuable objective, i.e. an objective that is worth pursuing for its own sake, without reference to other objectives. The philosophical basis of many objective functions is the view that it is intrinsically valuable to satisfy the preferences or improve the subjective experience of sentient beings (e.g., Broome 1991) – and all sentient beings, not just humans (Ng 1995). If an ASI is sentient, then its preferences or experiences arguably ought to be counted too. The question of whether an ASI can be sentient is very difficult to answer, touching on deep questions in the philosophy of mind (Chalmers 2010). If an ASI would be sentient, then its preferences or experiences potentially could be an important factor in a decision analysis.

Another complication for ASI decision analysis is the extremely high stakes. An AI could either solve a great many of society's problems or cause human extinction (Yudkowsky 2008). Either of these outcomes could dominate a typical decision analysis (Beckstead 2013). Published estimates of the value of preventing human extinction vary wildly, from $600 trillion (Posner 2004) to infinity (Weitzman 2009; Baum 2010). The value of solving society's problems could be at least as large.

One way to sidestep these complications is by using cost-effectiveness analysis (CEA). CEA seeks options that achieve some fixed objective for the lowest possible cost. For ASI risk management, a fixed objective could be avoiding ASI catastrophe. This objective can then be pursued regardless of how valuable it is. Likewise, regardless of how valuable is, society only has finite, limited resources available for reducing this risk. CEA can help identify how to allocate the resources to minimize ASI catastrophe risk.

To illustrate CEA of ASI risk management, consider the two options of relinquishment and Friendliness research. Major world governments may be able to pursue total relinquishment of ASI development. This might greatly reduce ASI catastrophe risks, but it could also be expensive to enforce and could have a large opportunity cost due to foregone benefits of ASI and related types of narrow AI. These costs reduce the merits of total relinquishment and could make governments less likely to pursue it. In comparison, ASI safety measures research may also reduce ASI catastrophe risks, and at much smaller cost than total relinquishment. Depending on the details, ASI safety measures research could be more cost-effective than total relinquishment.

5. Evaluating Opportunities for Future Research

Decision analysis can be of further help for guiding future research. The basic idea is that it is often helpful to design a research project in consideration of the decisions it can inform. From a decision perspective, research is of higher value when it better improves the performance of decisions by reducing decision uncertainty. High value research thus occurs when it brings decision makers new information that is relevant to the decisions at hand. Valuations of research can in turn inform decisions on the allocation of resources to various lines of research. Some prior research has considered the value of ASI risk research (Salamon 2009; Yudkowsky 2013 pp. 82-84).

The decision analysis concept Expected Value of Perfect Information (EVPI; Clemen and Reilly 2001 p. 512), can provide a formal quantitative approach to assessing the value of ASI

risk research. EVPI is the difference between the expected value of a decision with perfect information vs. with only currently available information. In the context of a decision analytic model, such as a fault tree or an event tree, the expected value of information is based on the extent to which information reduces the uncertainty about the value of a particular parameter in the model. Perfect information about a parameter eliminates that uncertainty.

In general, EVPI calculations are used to set an upper limit to how much should be spent on reducing uncertainty – research cannot produce better-than-perfect information. On their own, EVPI calculations cannot predict how valuable specific research will be in reducing uncertainty. However, even imperfect information can have great value in reducing decision model parameter uncertainties by some amount. Straightforward extensions of approaches to EVPI calculations can provide methods to assess the Expected Value of Imperfect Information Information (EVII; Clemen and Reilly 2001) and Expected Value of Including Uncertainty (EVIU; Morgan and Henrion 1990).

As with decision analysis in general, the expected value of information is typically evaluated using utility functions or functionally similar metrics. This introduces the same complications of evaluating the high stakes of ASI decisions. Barrett and Baum (2014) provide an approach to estimating value of information based on cost effectiveness that avoids these complications. This approach can be helpful for evaluating AI risk research.

6. Concluding Thoughts

We believe there is significant value in working towards a single integrated model that can represent the most important policy-relevant ASI catastrophe risk factors, and incorporate the best information available about those factors, all at a tractable level of detail. However, that model will not serve all purposes for all stakeholders. It is not tractable or useful to try to build a single model that tries to answer all potential questions, nor that tries to model all potential issues at an extremely high level of detail.

Even if an analysis cannot produce rigorous quantitative answers to decision questions, it can still be worth conducting. The risk analysis literature suggests that often, the most useful outcomes of a probabilistic risk analysis modeling effort are often not the model's outputs themselves (which may or may not be surprising to experts), but the new insights and improved communication regarding risks and risk management strategies that result from the structured thinking and multidisciplinary discussions needed for the analysis (Kumamoto and Henley 1996 pp. 132-136).

An important qualification that is important to recognize is that some aspects of ASI research could actually increase risks. The research thus should appropriately protect sensitive information, while providing description of methods sufficient to allow other researchers to examine and employ them. In general, we suggest protecting from general publication information that is non-obvious, not easily available from other sources, and that would be useful to actors with ill intent or with significant capacity for inadvertently causing harm. Similar rules of behavior for researchers are typically used in security-sensitive government work, and are being increasingly used in academic research.

In summary, risk analysis and decision analysis methods offer time-tested approaches to structuring assessment of risk issues to allow well-informed, transparent risk management decisions. This holds for emerging technology risks like ASI just as it does with other types of risk. The methods do not necessarily offer a "right" way of proceeding, and they are not purely a science; there is an art involved, with best practices suggested by empirical studies. However,

risk and decision analyses can serve several important purposes. One purpose is to help clarify key technical issues (both what is known and what could be learned with further research) in context of important decisions. Another purpose is to help separate technical issues from matters of value and policy debates. Both of these could be of great value for ASI R&D risks, given the range of stakeholders and potentially very high stakes involved.

Acknowledgements

References

Armstrong, S. and K. Sotala (2012). How we're predicting AI—or failing to. In: Beyond AI: Artificial Dreams. Pilsen, Czech Republic, University of West Bohemia: 52-75.

Armstrong, S., K. Sotala and S. S. Ó hÉigeartaigh (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. Journal of Experimental & Theoretical Artificial Intelligence.

Barrett, A. M. (2014). Analyzing Current and Future Catastrophic Risks from Emerging-Threat Technologies. Retrieved May 5, 2014, from http://research.create.usc.edu/cgi/viewcontent.cgi?article=1062&context=current_synopses.

Barrett, A. M. and S. D. Baum (2014). Value of GCR information: Cost effective reduction of global catastrophic risks (GCRs). Advances in Decision Analysis. Washington, D.C.

Barrett, A. M., S. D. Baum and K. R. Hostetler (2013). Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia. Science & Global Security 21(2): 106-133.

Baum, S. D. (2010). Is humanity doomed? Insights from astrobiology. Sustainability 2(2): 591-603.

Baum, S. D., B. Goertzel and T. G. Goertzel (2011). How long until human-level AI? Results from an expert assessment. Technological Forecasting & Social Change 78(1): 185-195.

Beckstead, N. (2013). On The Overwhelming Importance Of Shaping The Far Future. Department of Philosophy. New Brunswick, New Jersey, Rutgers University.

Bolstad, W. M. (2007). Introduction to Bayesian Statistics. Second ed. Hoboken, New Jersey, John Wiley & Sons.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. Journal of Evolution and Technology 9(1).

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford, Oxford University Press.

Broome, J. (1991). Utility. Economics and Philosophy 7(1-12).

Chalmers, D. (2010). The singularity: A philosophical analysis. Journal of Consciousness Studies 17: 7-65.

Clemen, R. T. and T. Reilly (2001). Making Hard Decisions. 2nd ed. Pacific Grove, California, Duxbury.

Cooke, R. (1991). Experts in Uncertainty: Opinion and Subjective Probability in Science. New York, Oxford University Press.

Dillon, R. L., C. H. Tinsley and W. J. Burns (2014). Near-misses and future disaster preparedness. Risk Analysis.

Fallenstein, B. (2013). Predicting AGI: What Can We Say When We Know So Little? Berkeley, California, Machine Intelligence Research Institute.

Joy, B. (2000). Why the future doesn't need us. Wired. Retrieved 9 October 2011, from http://www.wired.com/wired/archive/8.04/joy.html.

Keeney, R. and H. Raiffa (1976). Decisions with Multiple Objectives: Preferences and Value Trade-Offs. New York, John Wiley & Sons, Inc.

Keith, D. W. (1996). When is it appropriate to combine expert judgments? Climatic Change 33(139-144).

Kumamoto, H. and E. J. Henley (1996). Probabilistic Risk Assessment and Management for Engineers and Scientists. 2nd edition ed. New York, New York, IEEE Press.

Kurzweil, R. (2012). On Modis' "Why the singularity cannot happen". In: Singularity Hypotheses: A Scientific and Philosophical Assessment. A. H. Eden, J. H. Moor, J. H. Soraker and E. Steinhart. New York, Springer: 343-348.

Lempert, R. J., S. W. Popper and S. C. Bankes (2003). Shaping the Next One Hundred Years: New Methods for Quantitative, Long-Term Policy Analysis. Santa Monica, California, RAND.

Maher Jr., T. M. and S. D. Baum (2013). Adaptation to and recovery from global catastrophe. Sustainability 5(4): 1461-1479.

McGinnis, J. O. (2010). Accelerating AI. Northwestern University Law Review 104: 366-381.

Meyer, M. A. and J. M. Booker (1991). Eliciting and Analyzing Expert Judgment: A Practical Guide. London, Academic Press Limited.

Millett, S. M. and E. J. Honton (1991). A Manager's Guide to Technology Forecasts and Strategy Analysis Methods. Columbus, Ohio, Battelle Press.

Modis, T. (2012). Why the singularity cannot happen. In: Singularity Hypotheses: A Scientific and Philosophical Assessment. A. H. Eden, J. H. Moor, J. H. Soraker and E. Steinhart. New York, Springer: 311-340.

Morgan, M. G. and M. Henrion (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge, Cambridge University Press.

Muehlhauser, L. (2013). Mathematical proofs improve but don't guarantee security, safety, and friendliness. Retrieved May 10, 2014, from http://intelligence.org/2013/10/03/proofs/.

Ng, Y.-K. (1991). Should we be very cautious or extremely cautious on measures that may involve our destruction? Social Choice and Welfare 8: 79-88.

Ng, Y.-K. (1995). Towards welfare biology: evolutionary economics of animal consciousness and suffering. Biology and Philosophy 10: 255-285.

O'Hagan, A., C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow (2006). Uncertain Judgements: Eliciting Expert Probabilities. Chichester, West Sussex, England, John Wiley & Sons.

Omohundro, S. (2008). The basic AI drives. In: Proceedings of the First AGI Conference, Frontiers in Artificial Intelligence and Applications. P. Wang, B. Goertzel and S. Franklin, IOS Press. 171.

Posner, R. A. (2004). Catastrophe: Risk and Response. New York, Oxford University Press.

Pratt, J. W., H. Raiffa and R. Schlaifer (1995). Introduction to Statistical Decision Theory. Cambridge, Massachusetts, MIT Press.

Rayhawk, S., A. Salamon, T. McCabe, M. Anissimov and R. Nelson (2009a). Changing the frame of AI futurism: From storytelling to heavy-tailed, high-dimensional probability distributions. 7th European Conference on Computing and Philosophy (ECAP). Bellaterra, Spain.

Rayhawk, S., A. Salamon, T. McCabe, M. Anissimov and R. Nelson (2009b). The Uncertain Future. Retrieved 23 October 2011, from http://www.theuncertainfuture.com/.

Sagan, C. (1983). Nuclear war and climatic catastrophe: Some policy implications. Foreign Affairs 62(2): 257-292.

Salamon, A. (2009). How much it matters to know what matters: A back of the envelope calculation. Singularity Summit 2009. New York.

Sandberg, A. (2010). An overview of models of technological singularity. Workshop on Roadmaps to AGI and the future of AGI (at AGI10 conference). Lugano, Switzerland.

Sandberg, A. and N. Bostrom (2008). Whole Brain Emulation: A Roadmap. Future of Humanity Institute, Oxford University. Technical Report #2008-3.

Shulman, C. and S. Armstrong (2009). Arms control and intelligence explosions. 7th European Conference on Computing and Philosophy (ECAP). Bellatera, Spain.

Shulman, C. and A. Sandberg (2010). Implications of a software-limited singularity. ECAP10: VIII European Conference on Computing and Philosophy. Munich.

Slovic, P., B. Fischhoff and S. Lichtenstein (1979). Rating the risks. Environment 21(3): 14-20, 36-39.

Sotala, K. and R. V. Yampolskiy (2015). Responses to catastrophic AGI risk: A survey. Physica Scripta 90(1).

Stokes, J. (2008). Understanding Moore's Law. Retrieved January 25, 2014, from http://arstechnica.com/gadgets/2008/09/moore/.

Weitzman, M. L. (2009). On modeling and interpreting the economics of catastrophic climate change. Review of Economics and Statistics 91(1): 1-19.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In: Global Catastrophic Risks. N. Bostrom and M. M. Cirkovic. Oxford, Oxford University Press: 308-345.

Yudkowsky, E. (2013). Intelligence Explosion Microeconomics. Berkeley, California, Machine Intelligence Research Institute.