Forecasting Transformative AI

by

Richard Ross Gruetzemacher

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 8, 2020

Forecasting, Artificial Intelligence, Transformative AI

Approved by

David Paradice, Chair, Harbert Eminent Scholar, Department of Systems and Technology
Dianne Hall, Torchmark Professor, Department of Systems and Technology
Kang Bok Lee, EBSCO Associate Professor, Department of Systems and Technology
Kelly Rainer, George Phillips Privett Professor, Department of Systems and Technology

“The whole purpose of science is to find meaningful simplicity in the midst of disorderly complexity.” Herbert Simon (1991)

For my parents. Thank you for all of your support over the years.

Abstract

Technological forecasting is challenging, and the forecasting of AI progress is even more challenging. Forecasting transformative AI – AI that has great potential for societal transformation – which presents even more difficult challenges. This study addresses the research question of "How can we best forecast transformative AI?" To this end it explores a wide variety of techniques that are used for technological forecasting, demonstrates their viability in the context of transformative AI and evaluates their value for use in future efforts to forecast transformative AI.

The study focuses on scenario analysis techniques, a variety of judgmental forecasting techniques and simple statistical forecasting techniques. These include a survey, interviews, bibliometric analysis and the Delphi technique. The literature review identifies a new subclass of scenario planning techniques called scenario mapping techniques. These techniques are well-suited for forecasting transformative AI because of their incorporation of large numbers of scenarios (i.e., future technologies) in directed graphs. Two novel techniques that meet the criteria of scenario mapping techniques are developed and demonstrated.

The two methods, along with a variety of other forecasting techniques, are components of a proposed holistic forecasting framework for transformative AI. More generally, the holistic forecasting framework suggests that a combination of judgmental, statistical and scenario analysis techniques are necessary for forecasting complex future technologies such as transformative AI. This study is the first of which we are aware of to demonstrate the use of a holistic forecasting framework in any context. However, another framework(s) that satisfies the proposed criteria is identified and discussed.

The results of the study include forecasts generated from two of the techniques demonstrated along with significant insights for future work on forecasting transformative AI. A research agenda for forecasting AI progress was also created in demonstrating and evaluating the Delphi technique for the purpose of using expert elicitation to generate questions of interest for use as forecasting targets in prediction markets or forecasting tournaments. The results from the survey also generate insights into limitations of existing methods used for future of work research.

The study produces numerous novel contributions including theoretical elements from the definition of transformative AI, the holistic forecasting framework, the two novel methods and the concept of the subclass of scenario mapping methods. Beyond this, a large variety of insights are identified and reported as a result of the demonstration of the various techniques. Moreover, elicitation of some of the world’s leading experts on specific subdisciplines of AI yields some significant insights regarding future progress in AI.

Acknowledgments

I would like to thank Ashish Gupta for choosing to work with me, helping me to realize my potential for academic research and for helping me to put myself in a position to work on the world's most important problems. I would also like to thank David Paradice for letting me pursue my interests and for supporting me in a very non-traditional course of study. I would like to thank Shahar Avin, as well, for giving me an opportunity to collaborate on a much larger project that enabled me to further enhance the research conducted for this dissertation by strengthening my network and making available incredible opportunities that would have otherwise been out of reach. I also thank Kang Bok Lee, Dianne Hall and Kelly Rainer for agreeing to serve on my committee and for supporting this unique research topic. This section would be woefully incomplete if I failed to acknowledge the contributions of Abi Arabshahi, who, upon completion of my master's study on computational fluid dynamics encouraged me earnestly to pursue my passion (as he has continued to do since leaving, also), even against his own judgement, rather than continuing to work with him for my doctoral studies. Without the guidance and support of these academics at different stages of my 9 years of postgraduate study, this dissertation, and the body of research that has resulted from it, would not have been possible. Lastly, I would like to thank undergraduate instructors who initially helped me to realize my potential for post graduate studies: James Hiestand, Lucas Van der Merwe, Cecelia Wigal, Ron Bailey, Matt Matthews, Ron Goulet and Ed McMahon.

Regarding this study, I first would also like to thank all of the organizers of AI conferences where I collected data for consenting to me attending and soliciting participants for my study. This was a requirement for IRB approval (protocol #17-484), and I was only unable to get an affirmative response for one event. I also owe great thanks to all of the survey participants, interviewees,

workshop participants, Delphi participants and other experts who I spoke to over the course of this study. Without their willingness to engage with a PhD student from outside their own discipline for no publicity or recognition the completion of this expansive study would not have been possible.

Some of the work herein comes from prior publications, and some of this work was conducted with collaborators. Despite this, the major ideas presented here are my own. However, framing of these ideas and execution decisions for some of the elements of this study involved contributions from others. Particularly, in Chapter 2 Jess Whittlestone's contributions greatly enhanced the end result and I owe her great thanks. Also, in Chapter 7, and the other sections related to the Delphi's execution and analysis, I would like to thank my team, including Flo Dorner, Niko Bernaola-Alvarez and Charlie Giattino. Without their contributions the result would have been of lesser quality and would likely have been replaced with a poorly executed and low value Delphi demonstration in a less impactful context. I would also like to thank Danila Medvedev for helping to organize the scenario network mapping workshop in Moscow, and for his other numerous contributions in adapting the workshopping technique to be implemented using the NeyroKod concept mapping software package (also for granting me access to NeyroKod).

Finally, I have a list of fellow researchers in the AI research community, the AI strategy research community and the Effective Altruism community who (collectively) advised me on all different aspects of this study (order does not denote level of contribution): Jordi Bieger, Miles Brundage, Ryan Carey, Zoe Cremer, Eric Drexler, Adam Gleave, Ozzie Gooen, Danny Hernandez, Jose Hernandez-Orallo, Luke Muehlhauser, Alan Porter, Mark Ring and Jan Youtie. Because of the breadth of this study I have surely left out some parties who are due acknowledgement here, and if this is indeed the case, I apologize.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
DTAI	Dramatically Transformative Artificial Intelligence
FICT	Full-Inference-Cycle Tournament
FCM	Fuzzy Cognitive Map
GPU	Graphics Processing Unit
GPT	General Purpose Technology
HLAI	Human-level Artificial Intelligence
HLMI	High-level Machine Intelligence
JDM	Judgmental Distillation Mapping
NLP	Natural Language Processing
NLU	Natural Language Understanding
RTAI	Radically Transformative AI
SNM	Scenario Network Mapping
TPU	Tensor Processing Unit
TAI	Transformative Artificial Intelligence

1 INTRODUCTION

In a world of quick and dramatic change, forecasting future events is challenging. If this was not the case then meteorologists would be out of a job. Forecasting new technology is even more difficult because there is no past data to draw upon and the future technology itself is at best poorly understood (Roper et al. 2011). In the last decade significant milestones in artificial intelligence (AI) research were realized (Le et al. 2013, Silver et al. 2016, Radford et al. 2019, Vinyals et al. 2019), and the realization of these milestones left many to perceive the rate of progress in AI to be increasing[1]. This perceived increase in the rate of AI progress has led to substantial increases in investment, as well as increased public and governmental interest. Consequently, there has been a large amount of hype regarding AI progress, and a new research community has emerged for AI strategy[2].

A significant factor in maximizing good and positive impacts from new and emerging technologies is the anticipation of their possible effects. This requires assessing AI progress and forecasting future trends. Within the AI strategy community there is a growing group that is working to measure progress and develop forecasts for AI, with significant effort focusing on forecasting artificial general intelligence (AGI). In this study we focus on different forms of transformative AI (TAI; defined in detail in Chapter 2) because it is not yet clear how TAI will transform society. This notion includes both powerful narrow AI technologies (e.g.,

[1] As indicated by experts in interviews conducted for this study.

[2] Here, AI strategy is defined as the study of how to shape the impact of AI on society such that the maximum good comes from it.

comprehensive AI services; Drexler 2019) as well as progress toward systems capable of a broad range of tasks like AGI.

Despite the growing interest in the AI strategy community, very little work has been conducted that has made progress, either practically or theoretically, toward improving forecasts of AI progress or TAI. Moreover, the best techniques for traditional forecasting appear to be insufficient. Two of the world's foremost experts on political and organizational forecasting, Schoemaker and Tetlock, have noted that forecasting the next innovative technology from Silicon Valley is too difficult for even the best performing methods for geopolitical forecasting (Schoemaker and Tetlock 2016). In economics, trend extrapolation has been widely used for forecasting future events. However, as Brynjolfsson and Mitchell have noted, for forecasting the automation of jobs "simply extrapolating past trends will be misleading, and a new framework is needed" (Brynjolfsson and Mitchell 2017). This study explores and evaluates numerous forecasting techniques drawing from them to create novel forecasting techniques as well as a new holistic framework for forecasting TAI.

1.1 Artificial Intelligence

Given the many different aspects of AI it is important that we briefly introduce readers to the field. Although significant contributions were made earlier, the field of AI is thought to have formally begun when the name was coined for the 1956 Dartmouth workshop, the Dartmouth Summer Research Project for Artificial Intelligence (Russell and Norvig 2016). Since then AI has grown to encompass a wide variety of topics and sub disciplines such as automated planning, automated reasoning, computer vision, knowledge representation, machine learning, natural language processing (NLP) and search algorithms.

The motivation for this study comes in large part from advances in the subdisciplines of machine learning, computer vision and NLP. Machine learning concerns learning patterns from data that can be used to predict future phenomena associated with those patterns (Murphy 2012). Machine learning is commonly thought to be divided into three main types: supervised learning, unsupervised learning and reinforcement learning. Supervised learning requires labeled training data while unsupervised learning is used for unlabeled training data. Supervised learning is most often used for predicting future labels or future values based on the associations of values and labels with the data in the training set. Unsupervised learning involves recognizing patterns in data without having values or labels associated with the training data and its most common usage is for clustering. Reinforcement learning involves an agent interacting with its environment and learning from a reward signal[3], much like humans learn (Sutton and Barto 2018). In contrast to supervised and unsupervised learning, reinforcement learning can be thought of as a system generating its own training data.

The primary benefits of machine learning in recent years have been the AI subdisciplines of computer vision and natural language processing. Computer vision concerns perceiving objects through sensory data (Russel and Norvig 2016), and the current techniques used for this class of problems primarily involve machine learning. NLP is the study of techniques for processing speech and text (Manning and Schutze 1999), and the current techniques used here also are mostly

[3] Responding to a question about what reinforcement learning is during an interview conducted as part of this study, Richard Sutton, the lead author on the standard textbook on the topic described it: "Reinforcement learning is the idea, the ordinary idea that you can learn by interacting with the world. So, you don't have a teacher telling you what to do necessarily, but you do actions, some of them work out well for you and some of them work out poorly for you. And then you change the probability at which you do various things based on those things. So, in reinforcement learning we have to have a concept of goal. Of course, you always need a concept of goal of some kind, but in reinforcement learning it's very clear and explicit there's a number you get from the world … (that) you're trying to make it large. And that's called reward. You're trying to get the reward number high, so you assign that number and then we write our algorithms to interact with that number and prefer actions that are followed by high reward.

based in machine learning. Computer vision has benefited most from advances in supervised learning while NLP has benefited most from more recent advances in unsupervised learning.

The advances in the research areas described in the previous paragraphs primarily come from the emergence of deep learning, a type of machine learning that involves learning very complex representations of data hierarchically, with multiple layers of abstraction (LeCun et al. 2015, Schmidhuber 2015). Deep learning did not come about from new techniques – the techniques it uses have been around for more than 30 years (Rumelhart et al. 1985, Rumelhart et al. 1988, LeCun et al. 1989) – but rather from the decreases in computing costs that result from the increasing density of transistors in integrated circuits, described by Moore's Law (Moore 1965), and the use of graphics processing units (GPUs) for optimization. Deep learning relies on a class of machine learning models that involve large interconnected graphs, e.g., artificial neural networks[4]. Because of decreases in the cost of computation, larger models of this type with more layers[5] can be used. LeCun successfully used deep convolutional neural networks[6] as early as 1998 for the computer vision task of practically useful handwritten digit recognition (LeCun et al. 1998), but it was not until 2006 that the use of these methods was first described with the term deep (Hinton et al. 2006).

For several years deep learning was thought to be a powerful tool within a school of machine learning known as connectionism (Domingos 2015), but it otherwise was not thought of as any different from the majority of AI techniques that typically tend to excel on specific types of problems but do not generalize to broader classes of problems. In 2010, researchers at IDSIA[7]

[4] Or simply neural networks.

[5] The term *deep* is in reference to the large number of layers involved.

[6] LeCun's novel idea for convolutional neural networks was also crucial to the recent computer vision progress.

[7] IDSIA (Istituto Dalle Molle di Studi sull'Intelligenza Artificiale) is a non-profit Swiss AI lab.

began using novel techniques for recognizing handwritten digits that required the use of GPUs for optimizing very large neural networks (Cireşan et al. 2010). By utilizing GPUs, this research team realized that the techniques were not limited to handwritten digit recognition; they could be used for image classification (Ciresan et al. 2011). Their results were impressive within the community, but it was not until 2012, when a research group from the University of Toronto extended this work and used deep neural networks (DNNs) to win the annual ImageNET competition[8] (Deng et al. 2009, Krizhevsky et al. 2012) that the broader field of AI recognized the potential of deep learning.

Since then the field of AI research has changed dramatically. In 2013, DeepMind[9] used deep reinforcement learning[10] to outperform human experts on Atari video games using only the screen's pixel input[11] (Mnih et al. 2013). Around this time deep learning was also being applied to achieve breakthrough performance on a large number of tasks in NLP (Mikolov et al. 2013, Socher et al. 2013, Bahdanau et al. 2014, Pennington et al. 2014). Two years after the success with Atari, Google DeepMind went even further with a program called AlphaGo that used deep neural networks to beat a world champion in the ancient board game of Go (Silver, Huang et al. 2016). Go has significantly more decision possibilities than chess – i.e. it has a larger action space – and was thought to require a large degree of creativity and imagination for humans to excel, thus, to many in the AI community and those familiar with the game of Go this result was even more impressive than the Atari games result. Shortly after this success, deep neural networks began to

[8] Krizhevsky et al. won the ImageNET competition by such a large margin that each of the top ten placing teams the following year were using deep learning (Krizhevsky et al. were the only team using deep learning in 2012).

[9] DeepMind has since been acquired by Google and is now Google DeepMind.

[10] As mentioned previously, there are three types of machine learning: supervised learning, unsupervised learning and reinforcement learning. Deep reinforcement learning refers to the use of deep learning specifically for reinforcement learning. When used for supervised or unsupervised learning no distinction is typically made. Interested readers should refer to Russell and Norvig, 2015.

[11] Google DeepMind is now thought to be the world's leading AI firm. Their objective is to create AGI.

surpass human-level performance in the ImageNET competition (Deng et al. 2009, Szegedy et al. 2015, He et al. 2016). By 2017, Google DeepMind had developed a new deep reinforcement learning algorithm, AlphaGo Zero, that was able to beat their original AlphaGo 100 times to 0 without any expert knowledge[12] (Silver et al. 2017b); Google DeepMind then built a more advanced version, AlphaZero, that was able to become world champion in not only Go, but also chess and shogi (Silver et al. 2017a). Moving into 2018 and 2019 the breakthroughs did not slow: some of the most challenging Atari games have been conquered (Ecoffet et al. 2019, Ecoffet et al. 2020, Badia et al. 2020), one of the world's strongest professional video game players in the real-time strategy game of Starcraft II was defeated (Vinyals, Babuschkin et al. 2019), and OpenAI produced a language model that made incredible progress in NLP (Radford, Wu et al. 2019). This performance increase in NLP tasks was very impressive, setting substantial new benchmark records for the Winograd Schema Challenge and the Children's Book Test as well as generating blocks of text on common topics (e.g., recycling and unicorns) that were coherent and indistinguishable from human prose for passages of over 200 words.

The most recent advances in deep learning have been in NLP. These include the OpenAI model GPT-2 mentioned above and have all been a result of a 2017 breakthrough in deep learning that involved self-attention (Vaswani et al. 2017). The transformer architecture proposed by Vaswani et al. enables the exclusive use of feed forward neural networks to learn functions for sequential data (Yun et al. 2020), e.g. natural language data, better than any previous methods. In fact, the previous state of the art deep learning techniques that had been used for NLP, e.g. recurrent neural networks (Graves 2012) and long short-term memories (Hochreiter and Schmidhuber 1997), are now much less effective than transformer-based techniques for a number

[12] The original AlphaGo had used some limited expert knowledge to start the learning process.

of tasks. Furthermore, the performance gains from these models on tasks of general language understanding have been tremendous. However, not all of the breakthroughs in AI are related to deep learning. Because it enables humans to more naturally interface with machines it can be considered a major step toward more powerful AI systems (Sejnowski 2020). In fact, one of the most impressive results from last year did not use deep learning at all. A system, Pluribus, developed using reinforcement learning and game theory, was able to defeat professional human players in the six-player game of Texas hold-em poker (Brown and Sandholm 2019).

The previous paragraphs briefly describe the state and rate of AI progress, at least as it is commonly perceived. This study was designed to better understand what all of these impressive milestones actually mean about AI progress in general and what they indicate about future progress, both in the near-term and long-term. Given the breadth of the field and the highly technical nature of the field, this is a challenging task.

1.2 Artificial General Intelligence

Artificial general intelligence (AGI) is seen by many to be the grand goal of artificial intelligence research and it has the potential to transform society in ways that are difficult to anticipate. Not only are its impacts difficult to imagine, but the notion of AGI is ill-defined[13]; what may be indicative of general intelligence to some may not be sufficient for others, and there is no agreed upon test for general intelligence (Legg and Hutter 2007a). Others think human-level artificial intelligence (HLAI) to be the grand goal of AI. Those of this opinion typically believe in designing a more human-like intelligence (Minsky et al. 2004), though the difference between the two notions is not really clear. Again, there is ambiguity as to what constitutes HLAI; what may be

[13] Legg and Hutter have examined the issue closely and proposed a mathematically formalized definition for universal intelligence (Legg and Hutter 2007a, Legg and Hutter 2007b). However, through interviews conducted in this study, we have found that this definition is not widely accepted and that AGI and other similar notions are all commonly thought to be formally ill-defined terms by many who spend a lot of time thinking about them.

indicative of human-level intelligence to some may not be sufficient to others, and there is no definitive test for human-level intelligence[14]. There are other similar notions as well, such as high-level machine intelligence (HLMI) and strong AI. Different definitions are given in the literature (Searle 1980, Müller and Bostrom 2016, Grace et al. 2018), but all lack rigor and specificity[15]. For the remainder of this manuscript, we will refer to these terms collectively as either notions of AGI or simply AGI. Beyond notions of AGI is the idea of superintelligence (Bostrom 2014), which refers to a recursively self-improving AGI that surpasses human-level intelligence substantially, perhaps by an order of magnitude or more. This type of AGI potentially poses great danger if its values are not properly aligned with those of humans, and certainly raises philosophical issues, too. However, superintelligence is beyond the scope of this study.

The lack of an objective definition for AGI is due in part to the fact that we do not know how to create AGI. In fact, some believe that the primary goal of AI research is not AGI, but rather to learn more about our own intelligence[16]. Some further think that we will unlock an understanding of the mechanisms of our own intelligence with AGI. Regardless, we are uncertain how to create it. In theory, AGI could be instantiated by one algorithm (Hutter 2004) or constructed by a large number of different architectures with various components (Adams et al. 2012). Because of the ambiguity and ill-defined nature of AGI, it may be easier to forecast such systems by trying to anticipate their impact on labor displacement. Researchers concerned with forecasting AI and its economic impacts have begun to focus on the ability of systems to perform human tasks rather than entire jobs (Duckworth et al. 2019).

[14] Is child-like intelligence HLAI, or would adult intelligence be required? Or expert-level intelligence? Would adult-level intelligence in 90% of domains be HLAI? 95%? 99%?

[15] We propose a similar term that attempts to address this issue later in this study.

[16] This statement and the following statement are conclusions of interviews conducted as part of this study.

1.3 Transformative AI

Most AI researchers believe that AGI is something that will not come for 45 years or more (Grace et al. 2018), and thus not something we should think about too much right now. However, such a powerful AI technology has the potential to reshape socioeconomic structures and institutions, even systems of governance at the national and international level. Alternately, this kind of transformative change could be caused not by a single AI system, but by comprehensive AI services (CAIS) comprised of a large number of powerful AI systems, each capable of completing a type of or class of human tasks at the level of a typical human or above (Drexler 2019). AI technologies that are capable of this sort of societal transformation have been described by some scholars as transformative AI[17] (Dafoe 2018). The kind of transformation described has the potential to be comparable to or greater than historical periods of great transformation such as the agricultural revolution or the industrial revolution. Transformative societal change from advanced AI technologies may happen gradually or over a very short period of time. This uncertainty, and the general uncertainty of longer-term timelines for transformative societal change from advanced AI technologies are the motivation for this study. Like pandemics, many feel that global catastrophic risks from TAI are tail risks, and that they should thus be treated very seriously.

1.4 The Present Study

This study was designed to address the research question: "How can we best forecast transformative AI?" Thus, we seek to identify the most suitable techniques for forecasting TAI. This is accomplished by conducting a comprehensive literature review of forecasting techniques, which leads to a focus on methods that come primarily from the study of technology forecasting

[17] Dafoe defines TAI as "advanced AI that could lead to radical changes in welfare, wealth, or power" (Dafoe 2018). This definition is appropriate, however, as this study focuses heavily on TAI, a more thorough definition is required. The second Chapter of this dissertation is dedicated to defining TAI in more detail, and within the economic and technological forecasting bodies of literature on the topic.

and futures studies. The first layer of novelty in this study lies in the fact that it implements technology forecasting methods that have not yet been proposed or implemented in the context of AI forecasting. In fact, very little work has been conducted to assess the viability of different methods for forecasting AI progress, and this is the first study designed to do this rigorously. Through the course of exploring the existing techniques, we identify a gap in forecasting methods that seem most relevant to the topic of interest. In order to address this gap a novel technique is proposed and a proof-of-concept is presented for this new technique. To inform the new method the study also develops a novel workshopping technique for eliciting expert opinion. Finally, we go on to propose a novel forecasting framework for forecasting TAI. The new method becomes the central piece of the framework, and the workshopping process is a necessary input for the framework. We conclude by utilizing the Delphi technique to generate a research agenda for forecasting TAI.

The purpose of technology forecasting is to inform stakeholders and decision makers such that they are able to make the best strategic decisions for shaping and managing the impacts of emerging and disruptive technologies. A number of techniques are used to do this. In this study we begin by focusing on expert opinion, trend extrapolation and scenario planning. There are numerous ways to elicit expert opinion, and in this study we limit the focus to surveys, interviews, workshops and the Delphi technique. Similarly, there are numerous ways through which one can utilize expert opinion. We choose to compute forecasts directly from survey results. We compare these approaches with interviews using a mixed methods approach. We then use more interviews and workshops to inform scenario planning techniques. For trend extrapolation we explore the use of economic indicators and we also explore the use of bibliometric techniques for identifying trends in technology development.

In forecasting TAI we are attempting to inform those working on AI strategy research, and we are interested in the same goal: maximizing the good that results from the effects of AI technologies on society. Consequently, TAI forecasting is defined here as methodologically identifying near-to-mid-term[18] plausible future scenarios and, when possible, timelines that can inform AI researchers, government officials and policy makers for planning and making the best strategic decisions for maximizing future good from AI technologies. Dafoe has suggested the need for a mapping of the technical landscape in order to better understand the possibilities of and paths to TAI (Dafoe 2018). He also notes the potential role of scenario planning in this mapping. This dissertation addresses significant gaps in the literature regarding the utility and viability of various technology forecasting techniques for this mapping. A subclass of scenario planning techniques, dubbed scenario mapping techniques, is proposed. This class of techniques is very suitable for TAI forecasting. The novel technique developed in this study is a new example of scenario mapping techniques. The identification of the new class of techniques is another novel contribution of this dissertation.

The numerous novel contributions mentioned above are both practically and theoretically significant. As for the former, the implementation of these techniques will give the most rigorous and robust forecasts for TAI to date. These forecasts are of critical importance to those in the AI governance community as they are necessary for informing decisions about resource allocation and policy initiatives. As for the latter, the methods that are explored in this study have been previously used for organizational technology forecasting but have not been used in the broader context of forecasting the progress of a particular technology. The study builds on these methods

[18] We identify near-term as within five years, and mid-term as within five to fifteen years. We believe this to be a stretch given the lack of evidence for value of forecasts greater than five years into the future, but we note that forecasts do not have to assign timelines.

to contribute directly to theory in proposing a novel forecasting technique as well as a novel forecasting framework. These contributions will be useful for developing a direction for the future study of forecasting TAI.

1.5 Summary

In the previous pages a brief description of the advances in AI was given to explain the impetus for the significant progress in the field over the past decade. The chapter included a description of AGI, as well as other plausible TAI possibilities, and identified the motivation for this study. It also established the novelty and contributions to the existing body of knowledge.

The remainder of this study proceeds first with a careful analysis of the notion of TAI in the context of existing literature from economics, forecasting and technology forecasting. We then complete a comprehensive literature review of the forecasting literature relevant to technology forecasting and the methods that are relevant to this study. We then describe the specific methods that are used in this study. The novel new method and the novel forecasting framework are presented in the following chapter, followed by the results are presented from the forecasting techniques that were implemented during the course of the study, including a separate chapter for the results of the Delphi technique. This is followed by a chapter discussing the results and then by chapters detailing the contributions and limitations of the dissertation, respectively. This is then followed by a Chapter highlighting foresight and forecasts that resulted from the study, and the concluding comments.

2 DEFINING TRANSFORMATIVE AI

Transformative AI (TAI) is the focus of this dissertation. Consequently, there should be a rigorous definition of this notion because very little literature exists regarding it. In this chapter we first examine the little literature that does exist, as well as literature on technology forecasting that is relevant. We then define the notion of TAI by differentiating between TAI and dramatically transformative AI (DTAI). Then, another distinction is made between DTAI and radically transformative AI (RTAI). In the final section we discuss the implications of the AI/DTAI/RTAI categorization.

2.1 Background

We split this section into two subsections. In the first subsection we consider the existing notions of transformation from AI: from the varied uses of major stakeholders to the uses of members of the AI strategy community. In the second subsection, we examine literature from economics and on technological transformation.

2.1.1 Notions of Transformation from AI

Major management consulting firms, major accounting firms, leading think tanks and major governments have all recently begun discussing the transformative impacts of AI (McWaters 2018, PwC 2018, West and Allen 2018, White House 2019). For these groups, the notion of transformation is used relatively loosely; they do not all mean the same thing by it. Some focus on transformations to society while others focus on transformations in business. While more focus appears to be on the former, both academics (Fountaine et al. 2019) and industry experts (Ng 2019) have directed significant attention toward the latter. There is a great deal of variance in how the notion of transformation is used for all of these cases, much of it depending on framing and context. Moreover, many of those discussing these transformations from AI are discussing them in the present tense, believing AI to already be leading to transformations of society and business.

The term ‘transformative AI’ has recently begun to be used by philanthropic organizations (Karnofsky 2016) and members of the AI strategy community (Dafoe 2018, Zhang and Dafoe 2019) to refer to transformative change that is expected to come in the mid- to long-term future[19]. Thus, this notion of TAI relies on an understanding of transformative that is mostly inconsistent with how others are discussing anticipated transformations due to AI. The different definitions of TAI are shown in Table 1. These definitions are more in line with the original working definition of the notion that was intended when this study began. The final definition that is presented in this chapter is a compromise between how the majority perceive transformation from AI and how the AI strategy community has perceived TAI.

Table 1: Existing Definitions of Transformative AI

Karnofsky 2016	“potential future AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution”
Dafoe 2018	“advanced AI that could lead to radical changes in welfare, wealth or power”
Zhang & Dafoe 2019	“advanced AI systems whose long-term impacts may be as profound as the industrial revolution”

The definitions of TAI shown in Table 1 all refer to changes that are very radical. Dafoe (2018) explicitly mentions “radical changes” in the definition while the other two definitions

[19] Mid-term is defined here as five to fifteen years. Long-term refers to anything beyond that.

(Karnofsky 2016, Zhang and Dafoe 2019) invoke a reference to the Neolithic agricultural revolution and the industrial revolution. These revolutions caused unprecedented change for humanity: a transition from hunter-gatherer societies to settled civilizations, and a further transition from settled civilizations to the settlement of many large cities[20], staggering population growth and drastically increased life expectancies (North and Thomas 1977, van Neuss 2015). This discontinuous change in the metrics used to measure human progress, like life expectancy, has only been associated with the Neolithic agricultural and industrial revolutions (Muehlhauser 2015a). Consequently, many consider such transformation to be at a different level than that of other periods of human history (Drucker 1965, Bocquet-Appel 2011, Muehlhauser 2017).

TAI, when used in the cases described in the previous paragraph, is intended as an alternative formulation of notions of AGI or CAIS. Because we do not know what form(s) of AI will have transformative effects on society, speaking about transformative AI can be a more constructive means for discussing issues related to planning and resources allocation for mitigating risks associated with such transformation. However, many others are discussing the transformative impacts that AI is already having on society or transformative effects that are not as radical as those associated with the definitions above.

Some examples of major stakeholders discussing transformation from AI include Deloitte's concerns about how "artificial intelligence is transforming the financial ecosystem" (McWaters 2018) and the White House's assertion that AI is already transforming American life (White House 2019). Andrew Ng, the former Stanford professor and AI chief of Google and Baidu, has described AI as the new electricity, and anticipates it to transform how organizations operate in the coming decades (Ng 2019). In a similar vein, Fountaine et al. (2019) describe another

[20] More recently there has been stagnation in the growth of rural populations and exponential growth in cities (Bacci 2017).

transformation that businesses need to make to their operations in order to remain competitive in the age of AI. Others describe the anticipated societal transformation from AI in more drastic terms: a Brookings Institute's report describes the world being "on the cusp of revolutionizing many sectors through artificial intelligence" (West and Allen 2018) while PwC suggests that AI is set "to transform the way that we live and work" (PwC 2018).

2.1.2 Relevant Literature

Given the many references to the transformation that AI is causing or anticipated to cause in society and organizations, the radical notions of transformativeness expressed in the existing definitions of TAI seem to be extreme. In order to try to identify a more measured understanding of transformativeness in the context of AI, society and organizations we will review literature on long waves, technological transformations and general purpose technologies.

2.1.2.1 Long Waves and Technology

Long waves, or Kondratiev waves (Kondratiev 1926), are a macroeconomic principle that describes cycles of increased growth and slower growth. The cycles are thought to last between 50 and 60 years, and they are thought to include an upwave, associated with a period of prosperity, and a downwave, associated with a period of depression (Van Duijn 2013). The cause of long waves, and whether or not their existence can be explained by a single economic theory, are topics of academic debate. Thus, there has been substantial effort over the years to hypothesize theories of the causes of long waves because this could be very valuable for both economists and policy makers (Ayres 1990a).

Schumpeter drew from the work of Kondratiev and proposed that temporal clusters of innovations drive economic growth as well as instability (Schumpeter 1939). Consequently, his work is commonly associated with the idea that long waves are caused by innovation (or rather they are a byproduct of it). His work has significantly influenced all subsequent work on the impacts of innovation on macroeconomic phenomena as well as most work trying to explain the causes of Kondratiev's long waves. Mensch (1979) has proposed that recessions lead to an increase in major innovations due to the decrease in low-risk investment opportunities. He further contended that economic stagnation is a consequence of the lack of innovations to drive growth, but that these innovations will always reappear.

Much work has been aimed at developing theory for tying technological progress to long waves. Mensch et al. (1987) drew from a theory proposed by Nakicenovic (1987) about the historical replacement of old technology with new technology to formalize a theory for long-term economic cycles. Nakicenovic's theory featured the logistic substitution model and suggested that technological substitution should follow an S-shaped curve[21]. Another example of this is Mosekilde and Rasmussen's (1986) theory on technical economic succession and the long wave that draws upon Mench's idea of metamorphosis (Mensch 1979). Berry et al. (1993) offers yet another example of theory aimed at tying together technology progress and long waves. In this study, the authors draw from evidence from the US and UK economies to 54-year long waves to the diffusion and decline of techno-economic systems.

Theories tying technological innovation to long waves are also useful for the purpose of forecasting technologies. Goldstein has developed complex models using variables such as war, production, innovation, investment and wages to effectively predict a transition in the early 1990s from the stagnation phase of the long wave to the rebirth period (Goldstein 2006).

General purpose technologies (GPTs) are an idea from economics that refers to technologies that are characterized by being extremely pervasive in their applications across

[21] See section 3.2.2 for more on S-shaped curves and technology forecasting.

society, by spurring other technological development and by contributing to widespread productivity gains (Bresnahan and Trajtenberg 1995). Lipsey et al. (2005) define a GPT as "a technology that initially has much scope for improvement and eventually comes to be widely used, to have many uses, and to have many spillover effects." However, and contrary to other scholars, Lipsey (2007) explicitly plays down the importance of a productivity bonus when determining whether a technology can be classified as a GPT. Electricity is one of the most well known and uncontested examples of a GPT. It had a profound impact on society and led to changes in everyday life due to revolutionizing indoor lighting and communication media (Jonnes 2004).

2.1.2.2 Technological Transformations

Ayres tied together the theory developed by Kondratiev, Schumpeter and Mensch to propose a theory of technological transformations where recession leads to increased incentives in expectation for innovation (i.e., technological opportunity) that then lead to technological clusters that act as a catalyst for the upswings of long waves (Ayres 1990a, 1990b). While he does not explicitly discuss GPTs, Ayres does describe each of the five technological transformations that he identifies primarily in terms of the clusters of technologies which drove them. He also explains that the timing is mostly technologically determined. While this theory is strong, it still fails to adequately explain the cause for the downturns at the end of each period of prosperity.

Grinin et al. (2017) anticipated the idea of a cybernetic revolution. The cybernetic revolution they propose is a production revolution equivalent to the only other two production revolutions: the Neolithic agricultural revolution and the industrial revolution. These authors describe the cybernetic revolution as being characterized by the widespread use of self-regulating systems (the description appears to meet the criteria of what would constitute transformative change from AI; we discuss this further in a subsequent section). Similar ideas to technological transformation have been explored by others (Mosekilde and Rasmussen 1986), but are less relevant in this context.

2.1.2.3 Transforming General Purpose Technologies

The notion of a transforming GPT is relatively new, and one that is very relevant to the topic of consideration here. While Lipsey et al. (2005) makes clear the distinction between GPTs and transforming GPTs, most researchers consider the 24 transforming GPTs[22] proposed in their book to be examples of GPTs because it is the most comprehensive list, and they do not consider a distinction. However, the distinction made by Lipsey is directly relevant to the distinction necessary for different notions of transformation that are being used by major stakeholders and academics. Specifically, he states that transforming GPTs "lead to massive changes in many, sometimes most, characteristics of the economic, social, and political structures" while "other GPTs do not" (Lipsey 2007). The examples given by Lipsey of a GPT and a transforming GPT are lasers and electricity, respectively.

Table 2 depicts the five technological transformations of Ayres (1990a, 1990b), including brief descriptions, paired with the transforming GPTs (Lipsey et al. 2005), which were the drivers for each of Ayres' technological transformations. The steamship, motor vehicle, airplane and lean production are not included in Table 2 because they may be perceived to be redundant given the transforming GPTs that are reported in Table 2. This table is intended to more specifically demonstrate the relationship between GPTs and technological transformations as the definition of

[22] This list includes the domestication of plants, the domestication of animals, the smelting of ore, the wheel, writing, bronze working, iron working, the waterwheel, three-masted sailing, the printing press, the steam engine, the factory system, railroads, steamships, the internal combustion engine, motorized vehicles, airplanes, mass production, computers, lean production, the Internet, biotechnology and in the offing, nanotechnology. These technologies can each be classified into one of six categories: materials technologies, power, information and communications technologies, tools, transportation and organization.

transformative is also tied to GPTs. Thus, it is significant that the descriptions of Ayres' technological transformations correspond almost perfectly with the selected GPTs of Lipsey et al.

Table 2: Technological Transformations and Transforming GPTs

	Time	Technological Transformation	GPTs
1st	1770-1800	Change from water power to large-scale use of coal	Steam power
2nd	1825-1850	Steam power applied to textiles and railroads	Factories, railroads
3rd	1860-1900	Steel, mechanized manufacturing, illumination, telephones & motors	Electricity, internal combustion engine
4th	1930-1950	Advances in synthetic materials & electronics	Mass production
5th	1980-	The convergence of computers and telecommunications	Computer, the Internet

2.2 Unpacking TAI[23]

Transformative change is something that is difficult to objectively characterize. Regardless, it is important to do so for this study. To accomplish this characterization, we focus on indicators of transformativeness for determining whether a specific technology would meet the criteria. The most significant indicators are:

- Lock-in or irreversible change[24] in aspects of human life and flourishing. This refers to the idea that a technology or an application of a technology becomes so prevalent that it starts to exhibit characteristics of path dependence such that it becomes extremely difficult, or nearly impossible, to change paths (Shapiro and Varian 1998).

[23] The contents of this Chapter, including the following sections, are described in more depth by Gruetzemacher and Whittlestone (2019).

[24] By irreversible change, it is meant that the change is practically irreversible (i.e., that it is irreversible for all practical purposes).

- A productivity bonus associated with the new AI technology that strongly impacts many aspects of the economy with increased productivity. Widespread productivity gains have not yet been associated with AI, and this has become a topic of research for economists (Brynjolfsson et al. 2019).
- Anomalous patterns or discontinuities in the metrics used to measure human progress (e.g., life expectancy, the fertility rate or global gross domestic product).

The notion of lock-in used here is consistent with the idea of technological lock-in[25]. A common example of technological lock-in is the use of the QWERTY keyboard. It was first patented in 1878 and has been in use for over 140 years, despite being evidenced as a suboptimal arrangement of characters on the keys (Noyes 1983). However, because of the initial widespread adoption, the keyboard character arrangement is thought to be "locked-in" and unlikely to ever change[26] (David 1985).

Another example of technological lock-in is the use of nuclear weapons. Prior to nuclear weapons civilians did not need to immediately fear for their lives if global conflict were to begin, and governments did not have to fear the rapid escalation to an existential conflict. Following the use of nuclear weapons, the nature of great power conflicts was fundamentally altered, which led to the Cold War. Aside from the Cold War, there have been no direct great power conflicts since the use of nuclear weapons on Hiroshima and Nagasaki, but the world's superpowers all now must confront and avoid an existential threat that did not previously exist. Consequently, it is unclear as

[25] Although we are concerned with lock-in only in the simplest sense, without concern for the quality of the paths followed (Liebowitz and Margolis 1995).

[26] To change to a different arrangement it would take relabeling or replacing of existing English language keyboards throughout the world. On top of that, it would require retraining the entire computer proficient portion of the world's population to type using a new character arrangement. This would take a substantial amount of time (i.e., months at a minimum). Thus, the economic costs in lost productivity would be severe in the short-term, and the benefits from increased productivity might take a far greater time to be realized.

to whether or not the societal transformation caused by their use had a positive or negative effect. While nations still speak of and strive for nuclear disarmament, the Nash equilibrium of mutually assured destruction provides strong incentives for those who have nuclear capabilities to maintain the status quo.

The notion of lock-in or irreversible change is the foremost indicator of transformative change because it can apply to benign examples, such as QWERTY, or more extreme examples, such as nuclear weapons. On the other hand, anomalous patterns in the metrics associated with measuring human progress is primarily an indicator of dramatic or radical change. Historical examples of such patterns in these metrics can only be found in the Neolithic agricultural and industrial revolutions.

Transformativeness can also occur at many different levels in many different ways. It is easiest to think of the different ways in which technology can be transformative by thinking of different dimensions of transformativeness. The salient dimensions include:

- Extremity: this refers to the degree of the transformative societal impacts on the population impacted.
- Generality: this refers to the scope of the transformative impacts, whether they extend to society as a whole, or only to a specific domain, culture or nation.
- Fundamentality: this refers to the degree to which the transformative societal impacts change elements of basic life, with the most significant historical example being a change from nomadic living to civilized settlements during the Neolithic agricultural revolution.

These dimensions can also be thought to apply to other transformative technologies, such as Lipsey et al.’s (2005) transforming GPTs, or to production revolutions more broadly. Extremity is the most easily conceptualized of these dimensions, and perhaps the best suited for differentiating between the transformative societal impacts of different GPTs or AI technologies. Generality and fundamentality are also significant for more nuanced conversations about topics related to strategic planning and resources allocation for managing the societal transformation caused by TAI.

Given the indicators and dimensions proposed, we can now define TAI as "as any AI technology or application which leads to transformative societal change, that is, practically irreversible change to some important domain(s) or aspect(s) of society, potentially consisting of the lock-in of certain societal or technological trajectories" (Gruetzemacher and Whittlestone 2019). From this definition we can distinguish different levels of potential[27] societal transformation that can be thought to exist on the spectrum of extremity.

Based on this definition, we can make a further distinction between technologies that are associated with normal transformative change, dramatic transformative change and radical transformative change. Thus, there should be a clear distinction between TAI, dramatically transformative AI (DTAI) and radically transformative AI (RTAI). This can be thought of as the TAI/DTAI/RTAI categorization. We define these new levels of TAI below:

- TAI: AI technology that is capable of transforming society in a manner that is most closely analogous to GPTs, as defined by Lipsey (2007). Such technologies would certainly exhibit characteristics of lock-in, but would not lead to discontinuities in the metrics used to measure human progress. Societal change from such AI technologies could be expected to be analogous to societal change from bioinformatics[28] or nuclear weapons[29].

[27] Potential is critical here as we are discussing societal transformation in expectation. This is expounded upon in section 2.3.1.

[28] Lipsey et al. and Lipsey did not produce a list of GPTs, only transforming GPTs. So, there is no source to cite regarding GPTs which fall short of the requirements of transforming GPTs. Bioinformatics is at the intersection of two GPTs, and as a result, has been implied as a GPT in previous work (Appio et al. 2017).

[29] Nuclear energy is another debatable GPT, and perhaps even a transforming GPT. Ruttan (2006) has previously discussed nuclear energy as a GPT at length.

- DTAI: AI technology that is capable of transforming society in a manner that is most closely analogous to transforming GPTs, as defined by Lipsey (2007). AI technologies falling into this category could lead to discontinuities in the metrics used for assessing human progress, but these discontinuities would likely not be severe or applicable to the entire spectrum of these metrics. Such AI technologies could be expected to precipitate societal transformation analogous to that resulting from previous transforming GPTs such as the internal combustion engine or electricity.
- RTAI: AI technology that is capable of radical and profound societal change that is analogous only to previous production revolutions such as the Neolithic agricultural revolution and the industrial revolution. Possible societal transformation of this type would have no upper bound. Technologies precipitating this sort of transformative societal change would necessarily cause discontinuities in most or all of the metrics used for measuring human progress.

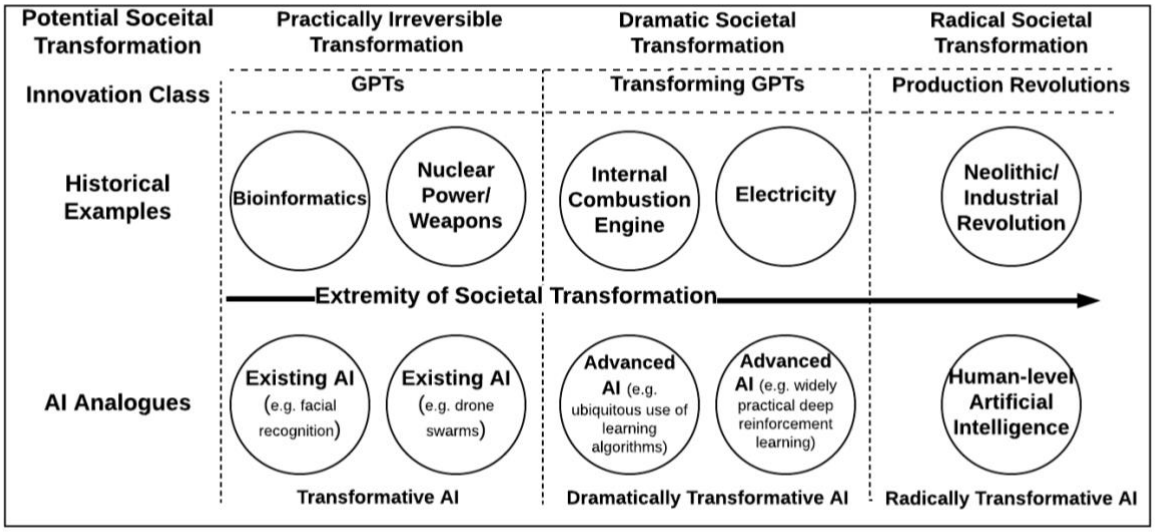

Potential Soceital Transformation
Practically Irreversible Transformation
Dramatic Societal Transformation
Radical Societal Transformation
Innovation Class
GPTs
Transforming GPTs
Production Revolutions
Historical Examples
Bioinformatics
Nuclear Power/ Weapons
Internal Combustion Engine
Electricity
Neolithic/ Industrial Revolution
Extremity of Societal Transformation
AI Analogues
Existing AI (e.g. facial recognition)
Existing AI (e.g. drone swarms)
Advanced AI (e.g. ubiquitous use of learning algorithms)
Advanced AI (e.g. widely practical deep reinforcement learning)
Human-level Artificial Intelligence
Transformative AI
Dramatically Transformative AI
Radically Transformative AI

Figure 1: This figure depicts historical examples of GPTs, transformative GPTs and production revolutions, each corresponding to levels of potential societal transformation: practically irreversible transformation, dramatic societal transformation and radical societal transformation, respectively. The historical examples are contrasted with hypothetical analogous AI technologies. The AI technologies corresponding to practically irreversible transformative change equivalent to GPTs are considered transformative AI; the AI technologies corresponding to transforming GPTs are considered dramatically transformative AI; and the AI technology that is analogous to the Neolithic agricultural and industrial revolutions is considered radically transformative AI.

Figure 1 presents two examples for each TAI and DTAI. Only a single example is given for RTAI because of the scarcity of historical analogues. The two examples for both TAI and DTAI are intended to represent potential different levels of impact. For the case of the TAI, the historical analogues used are bioinformatics and nuclear power. For the case of DTAI the historical analogues utilized are the internal combustion engine and electricity. We explore each of these cases further in the following paragraphs. These comparisons are made for the sake of demonstrating the utility of the proposed definition/framework, not because they have been explored at length[30].

For the case of nuclear energy, we will consider its demonstrated impacts rather than its potential impacts[31], and particularly focus on nuclear weapons. After their use on Hiroshima and Nagasaki during the second World War, the calculus of great power conflicts has dramatically changed. This has had effects on everyday life that were minimal for most, but likely moderate to severe effects on foreign relations. This also appears to be practically irreversible, because, despite efforts toward nuclear disarmament, the Nash equilibrium inherent in the notion of mutually assured destruction creates a strong path dependent trajectory. Thus, the effects of nuclear weapons are modestly extreme in expectation and minimally fundamental (in that they do not apparently reduce or increase the likelihood of war – outside of direct great power conflicts – they just transform its nature). They are also moderately general[32] because they impact less than half of the world's population directly. Similarly, the militarization of drone swarms has the potential to

[30] Each of these comparisons would likely require a dedicated paper to explore rigorously and robustly due to their highly nuanced nature and the fact that the potential transformation being discussed is in expectation.

[31] Nuclear power has the potential to be as transformative as the internal combustion engine once long-range space travel becomes a reality as it is the only practical source of energy (for both propulsion and life support) at distances far stars. See Lipsey et al. (2005) for further discussion of nuclear power as a potential GPT.

[32] Based on how the dimensions are defined for this framework for TAI.

transform the nature of war, and possibly increase the frequency of conflict, but it is unlikely to reduce the possibility of conflict. It could possibly increase the risk of nuclear exchange, but otherwise it is not likely to pose a threat to any significant portion of the world's civilian population. Consequently, it could be expected to lead to modest (extremity) societal transformation but would be minimally fundamental and moderately general[33] in that it would likely affect those in developed nations and nuclear armed states.

The impacts of bioinformatics may be difficult to assess and may in fact not yet be fully realized or even close to full realization[34], but widespread application with the single greatest impact is healthcare. It is very difficult to objectively assess the potential impact of bioinformatics on healthcare because of the numerous other technological advances in healthcare that will also continue to drive improvements in measures such as life expectancy. However, it would likely be considered modestly extreme in expectation. Other dimensions of its impacts would likely be highly general and moderately fundamental; highly general because new medical knowledge created from big data would not be limited to developed nations; moderately fundamental because life expectancy is a fundamental metric for measuring human progress, but the impacts would not extend to other such metrics like global GDP. Similarly, we can consider facial recognition. This would likely be modestly extreme because it would only affect privacy, and not implicitly induce substantial suffering or hardship; it would be highly general because it is a very powerful tool for law enforcement, and could be applied not only to urban dwellers, but all citizens registered in

[33] Considering the use of militarized microdrones or drone swarms for terrorism would necessitate these dimensions being adjusted. This underscores the importance of forecasting and scenarios in strategic planning and policy making.

[34] Bioinformatics relies heavily on advanced AI technologies and could be thought to be an AI technology. However, because it has been previously implied to have the properties of a GPT (Appio et al. 2017) we use it here as a historical example.

some form with the government. It would be moderately fundamental because it could fundamentally change government stability by removing the practical possibility of revolution[35].

We believe that the internal combustion engine is best classified as a transforming GPT. Here the impacts were not felt immediately but were ultimately highly general in that even those without vehicles can use public buses to travel throughout most of the civilized world. The impacts were certainly highly fundamental in that they had strong effects on global productivity (likely not discontinuous), created the new possibility of living and working at relatively far distances apart, and created a wide variety of powerful tools that could be used for a wide variety of tasks in both industry and in daily life. The effects were also moderately extreme, substantially changing the lives of those who were affected. Similarly, the ubiquitous use of learning algorithms could lead to the same levels of transformativeness[36] in all dimensions. However, we are now considering DTAI, and any further discussion of potential dramatically transformative effects would be no more than speculation. The methods we present in the later chapters of this dissertation are appropriate for exploring these notions; this example is simply intended to demonstrate the utility of this definition/framework for discussing forecasts and future scenarios involving transformative societal change from AI.

The impacts from the development of electricity are more dramatic than for the internal combustion engine. The transformative effects would be highly general, highly fundamental and highly extreme, thus, scoring highly in all of the dimensions. Widely practical deep reinforcement learning agents, able to learn from a small number of examples, when combined with the existing

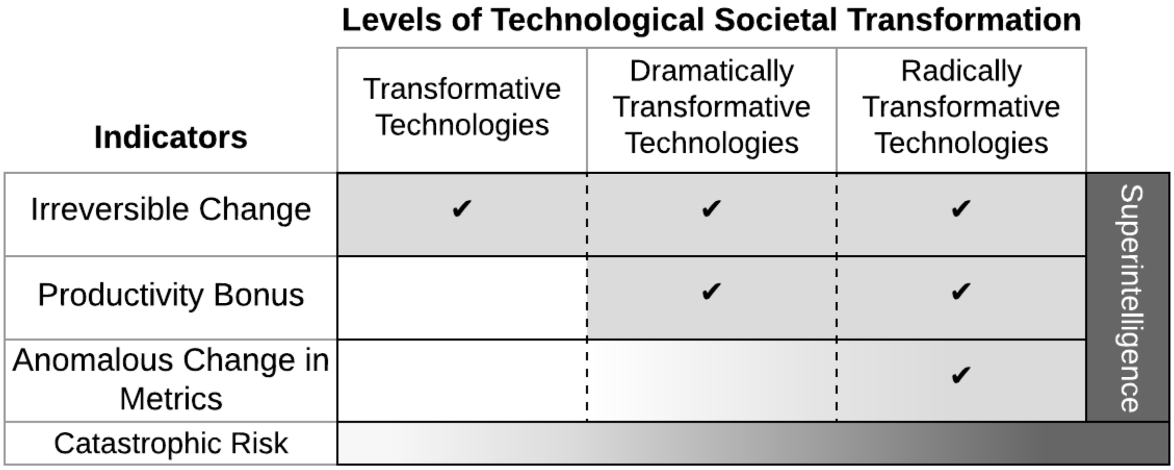
[35] The Arab spring can be thought of as a weak example. Tunisia and Libya did not utilize social media to suppress dissenters, while Egypt and Syria quickly imprisoned many revolutionaries who dissented.

[36] “The ubiquitous use of learning algorithms” is a poorly defined notion of advanced AI capabilities. However, it is likely that the use of existing algorithms, with no more fundamental breakthroughs, can be at a minimum as transformative as the internal combustion engine.

learning algorithms would be likely to generate transformative societal impacts to the degree of electricity. This is certainly debatable, thus the need for accurate forecasting and scenario generation, but similar to the previous DTAI comparison, any further analysis would be highly speculative.

The previous production revolutions – the Neolithic agricultural revolution and the industrial revolution – would be scored extremely general because they had unprecedentedly far reaching impacts. They would be scored extremely fundamental because for the majority of those who were impacted, they were impacted profoundly, as indicated by discontinuities in the change of metrics used for measuring human progress. They would also be seen to be maximally extreme because of the profundity of the impacts. Consequently, RTAI can be thought to score similarly on each of these dimensions. Anything short of that would likely still be considered DTAI.

In Figure 2 we return to the notion of indicators that was proposed earlier; namely the three indicators of irreversibility, productivity and anomalous changes in the metrics used to measure human progress. The above examples make it clearer how each of these different indicators is appropriate for the different levels associated with them. Specifically, lock-in is a necessary indicator for all levels of transformation from AI, but only sufficient for TAI. A productivity bonus marked by increased productivity across many sectors of the economy is a necessary indicator for DTAI and RTAI, but, when coupled with irreversibility, is only sufficient for confirming DTAI. Anomalous changes in the metrics used to measure human progress is a necessary condition for RTAI, although it can be present in mild forms for different manifestations of DTAI. These relationships are illustrated in Figure 2. The level of associated catastrophic risk is shown at the bottom and superintelligence (Bostrom 2014) is shown on the far right (for perspective).



Levels of Technological Societal Transformation

Indicators	Transformative Technologies	Dramatically Transformative Technologies	Radically Transformative Technologies	
Irreversible Change	✔	✔	✔	Superintelligence
Productivity Bonus		✔	✔	
Anomalous Change in Metrics			✔	
Catastrophic Risk				

Figure 2: this figure depicts a table which identifies the necessary and sufficient indicators for transformative technologies, dramatically transformative technologies and radically transformative technologies. It highlights the level of catastrophic risk associated with each of these as well as including a region identifying how superintelligence would fit for perspective.

2.3 Implications of the TAI/DTAI/RTAI Categorization

The definition of TAI proposed here is not straightforward like the previous definitions described earlier in this chapter. Defining TAI in the context of the existing literature on technological transformations, GPTs and long waves, requires three separate levels to effectively and constructively communicate. Here, we can first explain how this new understanding and the TAI/DTAI/RTAI categorization can enable more effective and productive dialogue between stakeholders trying to plan for safe and beneficial effects of TAI.

2.3.1 Benefits of the New Definition

The new definition is complex and requires separation into three distinct components. More than just a single definition, it provides a framework for talking about the effects of transformative AI at different levels. The distinction between TAI, DTAI and RTAI is significant in that it characterizes the transformative impacts of AI into three different levels. The first of these levels involves technologies that we are seeing today, and thus, some of the more impactful AI capabilities we see currently, are examples of TAI. Specifically, this includes the existing technologies that have the potential to become locked-in and consequently practically permanent features of our society, such as facial recognition technology. The second of these levels involves AI technologies, which may or may not exist presently, being used in ways that possibly create some (likely minor) discontinuities in the change of metrics used to measure human progress as well as numerous changes to trajectories of human development that are practically irreversible. This sort of change could be like that associated with the internal combustion engine or electricity. The third of these levels involves prospective AI technologies capable of radically transforming society, likely by automating most or virtually all existing economically valuable human tasks as well as creating some sort of discontinuities in the change of most or all of the metrics used to measure human progress. The only previous examples of this sort of change are associated with the two previous production revolutions: the Neolithic agricultural revolution and the industrial revolution.

The ability to speak about these three different levels of societal transformation has profound impacts on the abilities of organizations and governments to plan for the potential transformative impacts of AI. To date, much has been said about the potential of AI to transform different aspects of society, yet the various possible levels such transformation may take have been mostly ambiguous. The TAI/DTAI/RTAI categorization reduces this ambiguity by clearly presenting different levels in the context of historical technological analogues.

The distinction of RTAI is very significant. Notions of AGI or superintelligence have drawn attention, but there are too many negative consequences for using these terms while planning in organizations and governments; namely, that a substantial proportion of experts denounce them in principle and thus diminish the likelihood that they will be taken seriously by the major stakeholders[37]. By reframing this extreme case as RTAI, much of the nuance and negative connotations that exist with notions of AGI or superintelligence are likely removed[38]. Consequently, widespread use of this terminology could alter the public perception of the most extreme possible impacts of AI and could lead to organizations and policy makers taking more appropriate and realistic actions to prepare for these potential impacts and their risks.

The proposed dimensions offer another element of the framework which allows for conversations to not only focus on the extremity of the transformative change being discussed, but also to consider or focus on aspects of generality and fundamentality associated with the potential changes. The value of these elements of the definition/framework can be understood by considering their use in the end of Section 2.2 to dissect Figure 1, and to unpack the different historical technologies and their proposed AI analogues. These comparisons of juxtaposed technologies are much more easily, efficiently, and effectively conducted when using the different dimensions proposed in this definition for exploring their shared characteristics.

We can draw from the earlier comparisons to explore the potential transformative impacts from facial recognition technologies and autonomous drone swarms. At the end of Section 2.2 we described nuclear energy/weapons as being modestly extreme, minimally fundamental and moderately general. This makes for an easy comparison with drone swarms, an AI technology which would similarly have substantial, potentially transformative military applications. These two technologies can be easily understood each to transform the nature of great power conflicts. Based on the use of the definition/framework set forth here it may even be argued that due to the shared properties, drone swarms, like nuclear weapons after their first use, are likely to induce

[37] Their use often brings negative attention and invocation of Terminator-like scenarios.

[38] It is harder to say that these will lead to Terminator-like scenarios when by definition they are neither positive or negative and when notions of AGI or superintelligence already possess the negative associations with the Terminator-like scenarios.

rapid change toward a new Nash equilibrium (potentially less stable because it would likely be weaker than mutually assured destruction because civilians could be spared[39]) among great powers. This definition/framework enables easy discussion of the potential transformative effects of new AI technologies using historical analogies, thus making the discussions more concrete than simply describing them as transforming or transformative[40].

Another strength of the new framework for discussing transformative effects of AI is that it may be likely that TAI precedes DTAI, and DTAI precedes RTAI. This can help for the purposes of planning as well as for generating public support for the notions of dramatic and radical societal transformation. This actually opens the door to much more rigorous analysis on the likelihood of the lower levels of TAI preceding the higher levels.

One detail that is significant about the new definition/framework is that it refers to potential changes from transformative AI. More specifically, this means that the changes that are implied by the notions of transformativeness defined here and ascribed to the different levels of extremity are referring to the transformativeness in expectation. This is beneficial in that it enables the concept to be used theoretically and in practice before the outcome is realized. Also, because a certain technology may not be realized and thus may not have a significant impact, we should focus on numerous technologies with the potential for transformative societal change rather than just one.

Different technologies can be transformative because AI can power numerous GPTs; TAI can have many different manifestations (e.g., drone swarms and facial recognition can both coexist). Thus, these effects can be thought to exist on a spectrum, and the examples depicted in

[39] It could also lead to something else entirely if the first mover were to use it decisively while it still had technological superiority.

[40] Aside from the three defined uses of TAI, the notion of transformation with respect to AI has been used very ambiguously.

Figure 1 can be thought to be examples of different points on that spectrum (e.g., practical deep reinforcement learning assumes ubiquitous learning algorithms). This also implies that a critical mass of independent TAI technologies would lead to DTAI. After DTAI, increasing progress would increase the extremity and other dimensions of the impacts. Ultimately it would be expected that RTAI would be realized. However, all of these considerations are mentioned here to demonstrate the value of the new TAI/DTAI/RTAI categorization – these questions are all important and should be explored further with rigorous analysis.

Perhaps the most significant contribution of this definition/framework is the notion of DTAI. DTAI is significant for a number of reasons, not least because it is likely to be very transformative while also preceding RTAI. If DTAI preceded RTAI, it would likely have strong effects on the ability to ensure the safe and beneficial development of RTAI to benefit as much of the population as possible. However, while it would likely precede RTAI, it is not obvious that there is a level of societal transformation beyond the potential TAI anticipated from existing AI technology (e.g. drone swarms, facial recognition or a mild manifestation of ubiquitous learning algorithms). DTAI is also critically important because it represents technologies that present catastrophic and existential risks. Thus, those worried about risks of this kind should not focus solely on RTAI.

2.3.2 Connotations of the New Definitions

Based on the analogies provided in the descriptions of the different components of the TAI/DTAI/RTAI categorization, we can try to anticipate different levels of societal transformation – positive or negative – that could be expected with each of these levels of TAI. This is not intended to diminish the value of exploring different possible capabilities from advanced artificial intelligence, but rather, the new definition/framework proposed here is simply better suited for exploring different societal impacts as opposed to dissecting narrow AI technologies for specific capabilities. However, the framework could be used to analyze the specific capabilities of narrow AI systems with respect to their potential transformative impacts.

The definition/framework set forth in this chapter implies that transformative change be neither positive or negative. Erik Brynjolffson, a leading thinker on the economic impact of transformative technologies, has recently said (in the context of AI): "It could be the best 10 years ahead of us that we have ever had in human history or one of the worst, because we have more power than we have ever had before." (Gill 2020). For a simple example, there are numerous positive applications of AI that are being demonstrated in healthcare (McKinney et al 2020, De Fauw et al. 2018, Gruetzemacher et al. 2018), but here are also many potentially negative applications that may emerge, too (Agarwal 2019, Turchin 2018). This may lead to a balance between positive and negative impacts or may lead to a disparity either way. The ambiguity of whether societal transformation is positive or negative is implicit in this definition, and the definition/framework is intended to enable conversations about this issue to be more constructive and productive for ensuring that the next 10 years (as well as the next century and long-term human flourishing) be the best 10 years in human history rather than the worst. The proposed definition/framework here can be effective in facilitating better communication that can assist in meeting these goals because it enables more granular, efficient and effective dissection of the topics using the proposed dimensions of transformativeness.

2.3.3 Salient Implications for Current Dialogue

The new definition and TAI/DTAI/RTAI categorization can have potentially even more significant impacts. Sundar Pichai, the CEO of Alphabet, the world's largest AI companies and consistently among the world's five largest companies by market capitalization, recently made a statement at the World Economic Forum in Davos, Switzerland regarding the societal impact of AI:

> “AI is one of the most important things we’re working on ... as humanity. It’s more profound than fire or electricity or any of the bigger things we have worked on. It has tremendous positive sides to it, but, you know it has real negative consequences, [too].” -Sundar Pichai (Pichai and Schwab 2020)

Here, Pichai speaks of AI as being more profound than fire[41] or electricity. In the perspective just laid out of the TAI/DTAI/RTAI categorization, this would imply more profundity than highly general, highly fundamental and very extreme societal transformation; in other words, it implies that Alphabet is publicly – and at the world’s highest level gathering – anticipating RTAI. This struck the chairman of the World Economic Forum as salient, but the conversation quickly moved to near-term concerns, where it remained for the rest of the interview. Using language like DTAI or RTAI would enable such conversations to reduce the hyperbolic discounting tendency to focus on the near-term concerns[42], and could thus have a significant positive impact on mid- and long-term futures.

[41] Fire is not listed in Lipsey et al.’s (2005) list of 24 transforming GPTs, but there is no reason to think that it would not be considered a transforming GPT. Similar to electricity, it is a manifestation of a physical form of energy. It roughly characterizes the only practical means to extract energy from natural resources, and is necessary for many other transforming GPTs in the list such as the smelting of ore, bronze working, iron working and the steam engine. Due to this, it could be argued to be among the most significant transforming GPTs in the list.

[42] The near-term concerns are receiving enough attention, especially when the attention they receive is compared to the attention being afforded DTAI and RTAI. The attention being afforded RTAI is also higher than DTAI, and this could be problematic. This definition is also intended to bring attention to the largely ignored potential transformative impacts of DTAI.

3 LITERATURE REVIEW

While developing the definitions for TAI, DTAI and RTAI in the previous chapter, we extensively reviewed literature concerning theories behind innovation and technology development. It was suggested that these theories are also commonly used to develop models for predicting meta-level or long-term technological forecasts. In this chapter, we will revisit some of those elements, but will also explore a wide range of more common techniques that are applicable to technology forecasting. Specifically, this chapter focuses on judgmental forecasting techniques, simple extrapolative modeling, bibliometric techniques and scenario planning techniques.

3.1 AI Forecasting

The study of AI forecasting is in its nascency, and until recently, much of the work has relied on expert surveys. The oldest of these dates to a survey conducted in 1972 at a lecture series at the University College of London (Michie 1973). The majority of those polled indicated that they expected computing systems to exhibit intelligence at the adult human level by 2022[43]. Since 2006 twelve major surveys have been administered among experts and non-experts (Grace 2015, Zhang and Dafoe 2019). These surveys have been used to generate forecasts in the form of timelines. The most recent work has generated timelines by aggregating probability distributions fit to forecast quantiles elicited from participants (Grace et al. 2018, Zhang and Dafoe 2019). While the collection and aggregation of probability distributions from experts is an improvement upon previous studies on the topic, there remain many shortcomings in trying to quantify long-term

[43] These forecasts appear to be incorrect, but they cannot be evaluated until the end of 2022. Some experts contacted for this study do believe that these forecasts could be realized.

forecasts from expert opinions, the foremost perhaps being the questionable reliability of experts (Tetlock and Gardner 2016, Tetlock 2017).

The most rigorous of these expert surveys include four particular surveys which have been conducted since 2009, all pertaining to notions of AGI. The first was conducted at the AGI-09[44] conference and found that the majority of experts believed that HLAI would be realized around the middle of the 21st century or sooner (Baum et al. 2011). This study also found disagreement among experts concerning the risks involved with AGI and the order of certain milestones (different human-level cognitive tasks) leading to the development of AGI. These authors concluded that there was a reasonable chance that HLAI would be possible in the coming decades, and, because of the inherent risks associated with such advanced AI, this likelihood should be given more serious consideration.

The next of these studies consisted of a survey that was distributed among four groups of experts at the conference on Philosophy and Theory of AI in 2011, at the AGI-12 conference, to members of the Greek Association for Artificial Intelligence and to the top 100 authors in artificial intelligence by number of citations in May 2013 (Müller and Bostrom 2016). This was the first study of this topic to ask for experts' opinions through quantiles of 10%, 50% and 90% for the forecast target of interest, such as to elicit data that could be used to fit individual probability distributions. The authors of the survey questioned participants as to when they expected HLMI to be developed, and reported the experts to give a 50% chance of HLMI being developed between 2040 and 2050. These experts further indicated that they believed superintelligence would be created between 2 and 30 years after the emergence of HLMI, and that slightly over half believed

[44] The 2009 conference on Artificial General Intelligence.

that this would be a positive development – the remaining ~30% expected it to have negative consequences.

Yet another study was conducted by Walsh (2018). While this survey was ostensibly interested in "technological unemployment," conclusions regarding this were not specific or conclusive. The primary question of interest focused on HLMI[45], similar to that of Muller and Bostrom. Walsh surveyed AI experts, robotics experts and readers of a magazine article on the poker playing algorithm Libratus (Brown and Sandholm 2017). Expert forecasts for HLMI were not different at a statistically significant level for the two groups of experts, but the public opinion from the magazine readers was much more optimistic (by decades): the actual forecasts for the three groups were 2065, 2061 and 2039, respectively.

The most recent survey prior to the beginning of the current study was distributed to the primary authors of the 2015 Neural Information Processing Systems (NIPS[46]) conference and the 2015 International Conference on Machine Learning (ICML) (Grace et al. 2018). This study questioned participants on their forecasts of HLMI, but also included questions about a large number of specific tasks. All forecasters were asked to make forecasts at quantiles of 10%, 50% and 90% probabilities, which effectively elicited a probability distribution from each expert. This was not a new approach, but the analysis, including the aggregation of these probability distributions, was novel in the context of AI forecasting. The results indicated a median of 45 years until the development of HLMI, and a median of 120 years before all human jobs would be automated.

[45] For this reason this study is considered here and not with future of work related AI forecasting literature.
[46] From 2018 the acronym for the conference has been changed to NeurIPS.

The Grace et al. (2018) study was the most influential during the design of the survey instrument used as part of this study. This study significantly went beyond previous surveys, for example, the survey explored the framing effects of fixed years versus fixed probabilities questions. A small difference between the two framings was found, although the framing resulting in higher accuracy was uncertain. The primary question in the study was focused on forecasting HLMI, but this question was stated such that HLMI was defined as an AI system capable of performing all human tasks equal to or better than a typical human, for which the 50% median forecast was 45 years. However, when participants were asked to forecast when AI would replace humans at all jobs (the question was phrased such that the semantics were logically equivalent to the HLMI question using human tasks) the median 50% forecast was 120 years. This result seems to indicate some acute cognitive biases. The study also found Asian participants to have much earlier predictions than Europeans and North Americans.

A number of meta analyses of AI forecasting studies have also been conducted. In 2014 and 2015, Armstrong et al. and Armstrong and Sotala assessed previous timeline predictions that had been incorrect (Armstrong et al. 2014, Armstrong and Sotala 2015). They considered all known public forecasts for AI and proposed a decomposition schema for analyzing, judging and improving these previous predictions, applying it to five of the most notable forecasts. Muehlhauser has also conducted examinations of timelines and previous AI forecasts (Muehlhauser 2015b, Muehlhauser 2016). These two reviews offer the most comprehensive examination of HLAI timelines prior to those from the past decade. Regarding timelines, Muehlhauser concludes that we have learned very little from previous HLAI timelines other than the suggestion that it is likely we will create HLAI sometime in the 21st century. He further explores what we can learn from previous timelines and concludes with a list of ten suggestions[47] for further exploration of the existing literature.

AI Impacts[48] is a non-profit organization dedicated to improving our understanding of the likely impacts of HLAI. It is mentioned here because it has conducted significant work discussing techniques, curating related content and organizing previous efforts for forecasting HLAI, among other research and curation efforts that are less related to forecasting and more related to the mission of understanding HLAI impacts. It is typically thought of as an AI forecasting organization (some even consider it the preeminent organization in the space) despite its more general mission statement "to improve our understanding of the likely impacts of human-level artificial intelligence."

Recent work by Amodei and Hernandez presented a trend line for the increase in training costs for major milestones in AI progress between 2012 and 2018 (Amodei and Hernandez 2018). Specifically, the trend line revealed a positive correlation between training time and the achievement of AI milestones, with training time doubling every 3.5 months[49] for the milestones that have been realized between 2012 and 2018. However, several critiques of this have emerged (Carey 2018, Garfinkel 2018), the most compelling being that from a purely economic perspective the exponentially increasing rate for training costs is severely greater than the exponentially decreasing costs of computing. In contrast to all previous work in the literature on AI forecasting which has relied on expert opinion and extrapolation, the work in this study goes further by considering alternative forecasting techniques – namely scenario planning.

[47] His suggestions are limited in large part to further research on historical forecasts from individuals as well as how treatment of the subject has changed over the course of the academic study of AI.

[48] www.aiimpacts.org.

[49] The trend line is presented as an analogue to Moore's law for AI progress.

Dafoe discusses the notion of mapping technical possibilities, or the technical landscape, as a research cluster for understanding possible futures (Dafoe 2018). This notion is suggested as a requisite for a broader research agenda for AI governance. The proposals are not specific to HLAI, but rather are concerned with AI forecasting more broadly in the sense of TAI. Specifically, Dafoe advocates a multidisciplinary approach to mapping the technical landscape involving "expertise in AI, economic modeling, statistical analysis, technology forecasting and the history of technology, expert elicitation and aggregation, scenario planning, and neuroscience and evolution." The present study is the first to attempt this multidisciplinary mapping of the AI technical landscape proposed by Dafoe. Consequently, it contributes to the existing literature substantially. Dafoe also notes the important role of assessing AI progress and modeling AI progress, two topics that will also be discussed in this study. He separately discusses forecasting and its challenges, and also includes a desiderata for forecasting targets.

3.1.1 Future of Work

A topic closely related to AI forecasting is the study of the future of work. This refers to forecasting and foresight techniques which specifically focus on the economic impacts of new technologies on labor markets and is typically associated with economics rather than technology forecasting. However, it is relevant to the topics of this study, particularly because the major indicator of DTAI is a productivity bonus (i.e., economic productivity). Moreover, the economic impacts of AI progress are of significant concern to policy makers and should certainly be considered a topic related to AI forecasting more broadly.

Perhaps the most significant work on this topic was conducted by Frey and Osborne (2017). This study found that 47% of US occupations were at a high risk of computerization over the next two decades. This study also focused on jobs very granularly, with Frey and Osborne assessing

702 US occupations individually to arrive at their conclusions. This study utilized a novel technique that was based on data driven analysis. Other academic studies have similarly used a bibliometric approach (Das et al. 2020) while others yet have used surveys for eliciting expert opinion (Duckworth et al. 2019). Martinez-Plumed et al. (2020) utilize a novel, data-driven approach for mapping between AI capabilities and labor tasks using a cognitive abilities layer in the middle.

While Frey and Osborne were able to make projections related to specific jobs, computerization and information technology are more commonly thought to have an impact on some tasks comprising jobs more so than others (Autor et al. 2003). Moreover, AI is considered by economists to not only be a GPT, but to be a fundamentally new type of technology requiring a new framework for modeling as existing techniques such as trend extrapolation are ineffective (Brynjolfsson and Mitchell 2017). Economists also have pointed out that the effects of AI as a GPT are not yet being realized in the economy, a phenomena known as a productivity paradox, but the most likely explanation for this paradox is simply a lag between technology development and the realization of the economic potential of the technology (Brynjolfsson et al. 2019). Economists and future of work researchers typically look at task and job automation very granularly and do not model more extreme labor displacement scenarios that would suggest DTAI or RTAI. Moreover, these researchers suggest that notions of AGI are a long way away (Brynjolfsson et al. 2018) and let this suffice as justification for ignoring the effects of such technologies on labor markets. Frank et al. suggests even finer grained analysis for assisting decision makers and policy makers (Frank et al. 2019), which would preclude consideration of extreme scenarios which hinge on discontinuous technological development involving unknown unknowns that is not typically incorporated into econometrics models.

Many significant reports on the future of work have been compiled by think tanks and major business consultancies. While the methods used for these studies are not transparent, the conclusions all coalesce around a theme of there being significant labor displacement from automation and AI by the end of this decade (McKay et al. 2019, Harris et al. 2019). An earlier study from McKinsey found that "about half" of work activities are automatable given existing technologies (Manyika et al. 2017), and a more recent study from Brookings anticipates from 42% to 48% of paid working tasks for different states throughout the US to be susceptible to automation by the end of the decade (Muro et al. 2019). PricewaterhouseCoopers went further to make more specific predictions about jobs, suggesting that it would be possible for over 30% of jobs to be automated by the end of the decade (Hacksworth et al. 2018). These studies have significant weight in industry and in finance as their intended audience is managers and executives. However, there are many shortcomings in such industry reports that are uncommon in peer reviewed academic publications (e.g., point estimates of broad time horizons, loose language, etc.).

3.2 Technological Forecasting and Futures Studies

Technology forecasting is a broad field, and in this section we discuss select methods relevant to AI forecasting. A significant amount of literature on this topic comes from the study of technology management, and, consequently, many of the techniques are best suited for organizational management of technological innovation, strategic planning and similar applications. Futures studies is an academic discipline that is concerned with identifying and understanding possible, probable and preferable futures. Since AI forecasting is a nascent field of study, and one which is concerned with broader social good, we draw from both of these bodies of literature in order to develop the foundations for AI forecasting.

3.2.1 Technology Roadmapping

This is a widely used and flexible technique that is commonly used for strategic and long-term planning (Phaal et al. 2004). It is known to be particularly effective in structuring and streamlining the research and development process for organizations (Duin 2006). However, technology roadmapping can be used for planning at both an organizational level and a multi-organizational level. When applied, it often uses a structured and graphical technique that enables exploring and communicating future scenarios. It is generally thought to consist of three distinct phases: a preliminary phase, a roadmap construction phase and a follow-up phase (Garcia and Bray 1997).

Typically, there are eight recognized types of technology roadmaps including those for: a) product planning b) service/capability planning c) strategic planning d) long-range planning e) knowledge asset planning f) program planning g) process planning and h) integration planning (Phaal, Farrukh et al. 2004). Furthermore, there are eight types of graphical formats for roadmapping: a) multiple layers b) bars c) tables d) graphs e) pictorial representations f) flow charts g) single layer and h) text. The broad range of types and graphical formats, as well as the core three phases of roadmapping, make the method flexible. In application, these techniques are frequently molded to fit the specific case.

The Association for the Advancement of Artificial Intelligence and the Computing Community Consortium have recently completed a 20-year Roadmap for AI research (Gil and Selman 2019). The first phase focused development on three paths (integrated intelligence, meaningful interaction, and self-aware learning), each developed in a separate workshop. The second phase was the generation of the roadmap and the third phase was getting feedback from the community for revisions. This example can be thought of as a combination of the different types of roadmaps that relies on a combination of the different types of graphical formats for reporting, thus illustrating the flexibility of the method. However, technology roadmapping techniques have limitations, such as their heavy reliance on visual modeling (Roper et al. 2011). Technology roadmapping techniques have also failed to produce acceptable results in the previous attempt to use them for developing a roadmap to AGI (Adams et al. 2012, Goertzel 2014b, Goertzel 2016).

3.2.2 Extrapolation

Extrapolation from past events to the future is perhaps the most intuitive approach to forecasting. Although frequently referred to as naive forecasting, it is a powerful approach, yet also inherently dangerous and easily misused (Roper et al. 2011). Despite the dangers, the technique is among the most valuable, reliable, objective and effective forecasting methods when used for many practical purposes (Armstrong 2001). Extrapolation is commonly used for forecasts of up to two years and is used for longer-term planning in some cases. While very powerful, it is still but one of the many techniques in a forecaster’s tool belt that are necessary for creating forecasts which are reliable and robust for planning and strategic decision-making.

Extrapolations can be made for time series data as well as for cross-sectional data (Armstrong 2001). For time series data it is assumed that all necessary information is contained in the historical values. For cross-sectional extrapolation it is assumed that evidence from one dataset can be generalized to another. In either case, the extrapolations rely on variables that are aggregate measures of phenomena which are somehow related to the technology. Such variables are selected because we believe them to be good indicators of future progress. In technology forecasting terminology these are known simply as indicators (Roper et al. 2011). Useful indicators are typically considered to be one of two different types: technological indicators or social indicators (including economic indicators). A more extensive discussion of good indicators can be found in the following subsubsection (as well as in Porter and Cunningham 2004).

The logistic substitution model formally theorizes the S-shaped curve notion that had previously been used for many types of technological forecasting (Fischer and Pry 1971, Nakicenovic 1979, Twiss 1986). (However, not all growth curves take the form of a logistic function (Martino 1993, Kucharavy and De Guio 2008).) The S-shaped curve has particular value for technology forecasting and is commonly used with extrapolation to model technology development and innovation. The S-shaped curve can be best conceptualized when breaking the curve into three sections. The first, at the bottom includes the development of the technology, the proof of concept, the demonstration of different practical applications, etc. This period can sometimes take decades before the technology becomes widely used in very economically practical applications, which is what can be considered the second part of the curve. Because it can take a while for development and application to practical applications, the first period of an S-curve could be considered an explanation for the lag that is being observed for AI to increase productivity (Brynjolfsson et al. 2019). The third and final phase is a plateau, where, in many cases, a new technology sets in and the exponential increase in productivity continues by successive S-curves. A recent example of this is battery energy density, which has continued to double from 1985 to 2005 over the life cycles of three different battery technologies: nickel-cadmium, nickel-metal hydride and lithium-ion. Figure 3 depicts stacked S-curves in an ideal case.

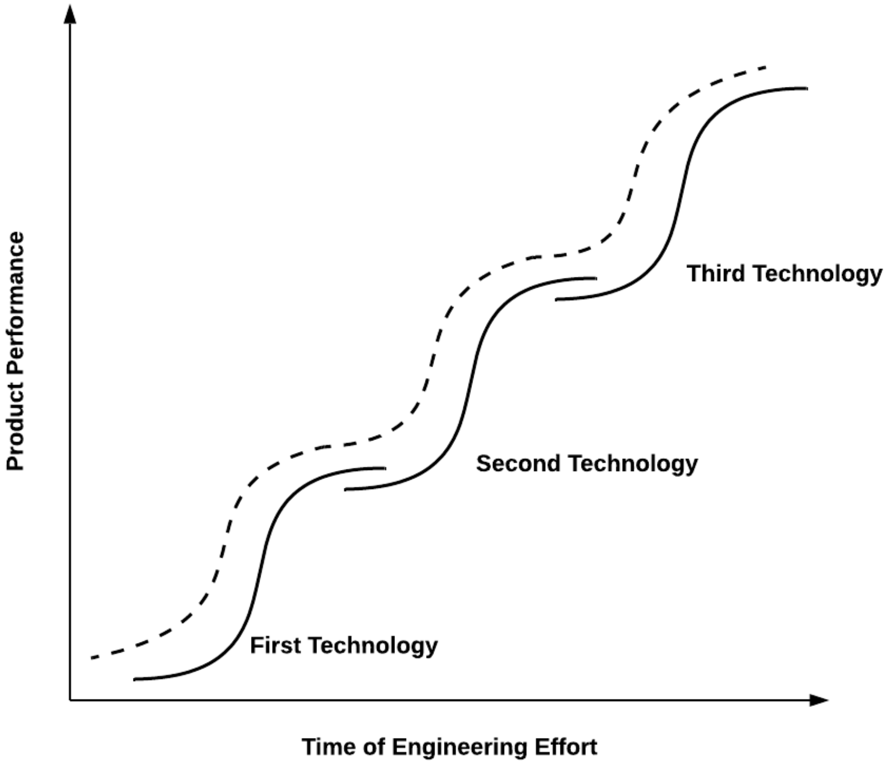


Figure 3: Stacked S-curves in an ideal case.

Extrapolation has had much success for many different types of forecasts. In fact, it has been used for one of the most successful attempts to forecast progress in AI. In 1994, Russel and Norvig (1995) plotted the performance of AI systems at chess using ELO. The exponential curve they fitted to the data accurately predicted that these systems would reach the level of world champion three years later, in 1997, when IBM's DeepBlue defeated Gary Kasparov. Extrapolation has also had significant practical application in a field closely related to progress in computing and AI: semiconductor development. It is well known that Gordon Moore worked at Fairchild Semiconductor in the 1960s. During this time, Moore collected data about many different metrics for measuring semiconductor progress, and upon comparing all of these metrics he found that the rate of increase of transistors per chip fit a nice curve. This curve is now widely known as

Moore’s law, and this model became regarded so well that it ultimately began to be used to dictate progress when it was incorporated into the Semiconductor Industry Association’s industry roadmaps (this is also an example of how, when paired with strong forecasts, roadmaps can be practically useful and can dictate the progress of an industry). Moore’s law no longer appears to be valid, but a generalized Moore’s law which suggests that the number of computations per dollar doubles in 18-24 month intervals still roughly holds.

3.2.2.1 Indicators

As noted, there are two basic types of indicators that are of interest in technology forecasting: science and technology indicators and social indicators (Roper et al. 2011). Science and technology indicators, or simply technology indicators, are directly related to the progress of the technology of interest. Social indicators are intended to collectively represent the state of a society or some subset of it. Technology indicators ideally must adhere to three restrictions: 1) the indicator must measure the level of a technology’s functionality, 2) the indicator must be applicable to both the new technology and to any older technologies it replaces and 3) there must be a sufficient amount of data available to compute historical values. In reality, many times indicators are not available which satisfy all of these requirements. In such cases efforts should be made to identify indicators that suffice as best as is possible. Social indicators can include economic factors, demographic factors, educational factors, etc., and they are thought to be analogous to a technology’s functional capacity. Predictive models for forecasts can be created using either a single indicator or through multiple indicators.

3.2.2.2 Indicators for Assessing AI Progress

Assessing progress in the development of technologies is an essential element to forecasting their future progress because aggregate measures for this are typically appropriate technology indicators. Assessing progress in the development of AI technologies is especially difficult because AI technologies are used for a wide variety of human tasks in which their performance is measured with respect to the level of a typical human. Thus, it is important to identify the indicators that satisfy the three requirements set forth earlier as closely as possible. Efforts have been made, such as those by Amodei and Hernandez (2018), but criticism of their work illustrates the difficulties (Carey 2018, Garfinkel 2018).

As there are a wide variety of tasks that AI is designed to perform, there are a wide range of benchmarks for progress in narrow domains which can serve as indicators. However, these are poor indicators of general progress; superior indicators of general progress are much more challenging benchmarks than the indicators for narrow domains[50] (Hernández-Orallo 2017). Thus, benchmarks for AI progress that are useful today will not remain useful once AI technologies surpass the human standard for the specific domain of the benchmark. Hernandez-Orallo has conducted a thorough review of the many previous efforts at AI evaluation that offers some guidelines for future attempts at AI evaluation (Hernández-Orallo 2014). A desiderata or similar guidelines created specifically for AI forecasting indicators[51] could be very useful, also.

Efforts to identify variables as indicators of broader AI progress are not as extensive as the previous work on surveys of expert forecasts, but they have been ongoing for a number of years (Grace 2013). Other work has attempted to address the challenges of assessing AI progress and building models for forecasting future progress based on AI progress indicators. Muehlhauser and Sinick (2014) proposed three core inputs for modeling AI progress: 1) AI funding 2) quality-adjusted researcher years (QARYs) and 3) computing power. Brundage[52] drew on suggestions of

[50] The Russell and Norvig use of ELO score for anticipating AI progress in chess is an example of success in a narrow domain, but no examples have been successfully applied for broader purposes in the context of AI.

[51] Perhaps similar to the desiderata for AI forecasting targets proposed by Dafoe (2018).

[52] Brundage gave a brief review of technology forecasting literature, including three references.

Hernandez-Orallo (2014) for hierarchically modeling progress using Carrol's hierarchical model of intelligence (Carroll 1993) to notions of AGI to propose an outline for research toward developing a rigorous framework for modeling AI progress (Brundage 2016). Martinez-Plumed has most recently led efforts to explore the different dimensions of relevant indicators (Martinez-Plumed et al. 2018) as well as frameworks developed from Item Response Theory (IRT; Martinez-Plumed and Hernandez-Orallo 2019).

Performance on human tasks can be thought of as a social indicator for assessing the impact of artificial intelligence on labor markets (Duckworth et al. 2019). While such studies have not considered the use of these same indicators for modeling progress toward more powerful systems, such as DTAI or RTAI systems, there is potential to combine labor displacement studies and more general forecasts regarding AI progress. At the least, valuable data from future of work projections could increase signal in very broad forecasts which utilize a large variety of different methods. Jobs could also be used as a social indicator for these purposes. As noted previously, Frey and Osborne notably assessed the automatability of 702 professions using a novel method, concluding that 47% of the professions were at high-risk of automation (Frey and Osborne 2017). They further found social intelligence, creative intelligence, perception and manipulation tasks to be bottlenecks to automation over the next two decades. However, tasks are now thought to be more appropriate for forecasts because AI technologies may replace some tasks comprising a job and not others (Autor et al. 2003). Here, tasks would be simply defined as work activity units which produce output (Autor 2013). Using tasks as an indicator can be challenging also because tasks are multidisciplinary and can be difficult to define. To address these issues, Fernandez-Macias et al.

proposed definitions of tasks from two primary perspectives: socio-economic and computational[53] (Fernández-Macías et al. 2018). This led to a proposed framework for assessing the impact of AI on the future of work considering two major features of tasks – autonomy and generality – which are central to this assessment. While this work is a significant constructive contribution, and it is unclear whether this framework could be adapted for use as social indicators, this is a promising direction for future research.

3.2.2.3 Tech Mining

"Tech mining" refers to a broad set of techniques that can be used to generate indicators from big data sources. Porter and Cunningham discuss the use of innovation indicators for understanding emerging technologies, and propose nearly 200 such indicators (Roper et al. 2011). Many of the proposed indicators are relevant to organizational management, but some offer insight into possibilities of indicators that can be used for assessing and forecasting AI progress (Porter and Cunningham 2004). AI Index creates an annual report each year that monitors a range of indicators of AI progress, such as conference attendance, benchmark performance and patents filed (Perrault 2019, Bauer 2018). However, these indicators tend to fall short of those typically associated with tech mining in that they only report descriptive statistics rather than using big data analytics techniques. Common techniques in tech mining utilize data from numerous sources including news, patents and academic literature. Such techniques are known as patent mining, bibliometrics and scientometrics. Tech mining is widely used in technology forecasting applications related to innovation management and has begun to see use recently for mapping landscapes relevant to AI research (Omar et al. 2017, Zhang et al. 2019).

[53] "A socio-economic perspective as considered in the social science literature on the work process [see e.g. Braverman, H., Labor and Monopoly Capital 1974] and a computational perspective in terms of how the task is seen from the point of view of computerization."

Recent efforts have been made to apply tech mining to the assessment of AI progress, but these methods are still in early stages of development. Martinez-Plumed et al. (2020b) proposed the AIcollaboratory, an open-source project for a data-drive framework that is intended to enable researchers to collect and explore data on AI progress, results and capabilities[54]. This project focuses on AI capabilities and attempts to determine how AI technologies are able to generalize across differing tasks. Barredo et al. (2020) also use tech mining to evaluate the research communities behind the progress in over twenty leading AI benchmarks, ultimately identifying numerous features and links associated with different groups in the communities. Each of these are good examples of the value that can be offered by tech mining. Because of the power of tech mining, the largest current effort in assessing AI progress and AI forecasting is currently focused on collecting data for mining.

3.2.3 Expert Elicitation

Forecasters and stakeholders rarely have the technical expertise and broad understanding of the many dimensions of technologies being forecast to create accurate forecasts (Roper et al. 2011). However, experts commonly do have the necessary expertise and understanding as well as tacit knowledge that can be valuable to forecasters in creating forecasts. For these reasons expert opinion is frequently sought when developing forecasts of future technologies. Expert opinion can be used directly to develop forecasts or scenarios, and it can also be used for approving or validating models that were developed from indicators without expert input (Armstrong 2001a, Porter and Cunningham 2004).

The formalized, methodical and documented procedure for obtaining and combining probabilistic judgements is known as expert elicitation (Colson and Cooke 2018). Expert

[54] This project can be found at https://www.github.com/nandomp/AICollaboratory.

elicitation can be conducted in a variety of ways and even the process of selecting 'experts' can be challenging. Thus, it is critical to carefully select how and who to use when eliciting expert opinion.

The question of how to elicit expert opinion has a wide variety of answers. Beard et al. (2020), in reviewing techniques of probabilistic forecasting for existential risk assessment, identify three primary categories of expert elicitation techniques: individual expert elicitation, group expert elicitation and structured expert elicitation. We will consider these categories of expert elicitation techniques as well.

3.2.3.1 Surveys

Surveys solicit expert opinion from multiple experts without interaction between them. This technique is widely used because it is straightforward to implement and relatively inexpensive (Roper et al. 2011). Challenges to this method include sampling difficulties, especially those due to nonresponses and sample biases that occur when the sample poorly reflects the desired or optimal population. During the development process, after careful selection of the questions to include, it is essential to pilot the survey with a small group that is roughly representative of the population of interest. Feedback from this exercise can reduce significant errors that could pose problems during analysis. An extensive discussion of survey methods, from sampling and question design to analysis and survey interviewing techniques, is given by Fowler (Fowler 2013).

The Cooke method (or the classic method) of assessing the quality of expert judgements for expert elicitation comes from the field of risk analysis (Cooke 1991, Morgan 2014). It is a very powerful technique for structured expert elicitation, and it can be used for group expert elicitation like surveys (Beard et al. 2020). This technique involves the inclusion of calibration questions to calibrate the experts' forecasts so that they may be weighted during aggregation (Aspinall 2010).

Aggregated forecasts using weights derived from calibration questions or previous forecasts have been shown to be more accurate than individual forecasts (Tetlock and Gardner 2016). Furthermore, calibration training can improve performance on forecasts and aid weighting (Moore et al. 2017).

Surveys can be found across numerous applications in technology forecasting and often fall into one of two classes for data collection: either by mailed or emailed surveys or through in person interviews. Mailed or emailed surveys are more simply implemented (Roper et al. 2011), but in person interviews can provide richer qualitative data (Creswell 2014). Mailed or emailed surveys are not only for quick elicitation of opinion from large groups but can also be used for eliciting detailed information from experts (Curtright et al. 2008). Thus, expert surveys are commonly used with small groups of experts (Wise et al. 2016). Nemet et al. (2017) have gone further to identify a small subset of "leading-experts" from among participants of a larger sample of 166 experts. This suggests that the expertise of select experts is more valued than that of experts randomly sampled.

Interviews are most commonly used in the context of technology forecasting as an alternative technique to collecting data in surveys. In the social sciences, when used in this manner, they are more frequently referred to as structured interviews and enable richer data collection due to the ability to ask unique follow-up questions for clarification and elaboration to subjects' responses (Creswell 2014). Tailored methods for such studies have been very successful using extremely small groups of experts (i.e., 3-7; Baker et al. 2009, Baker et al. 2010, Baker et al. 2011), and can be used for eliciting probability distributions by surveying participants on complex models and aggregating the results (Clemen and Winkler 1999) rather than by eliciting probabilities in the form of quantiles directly. However, structured interviews are not necessarily limited to data collection on small scales (Kassie et al. 2013). Such surveys are not necessarily limited to use for qualitative purposes, either (Nakagawa et al. 2010), and they can have other added benefits such as the potential for reducing experts' (or participants') overconfidence due to the more personal form of data collection (Bistline 2014).

3.2.3.2 The Delphi Technique

The Delphi technique is an alternate method of structured expert elicitation that was developed at the RAND Corporation in the 1950s in tandem with the development of scenario planning methods (Bradfield et al. 2005). This approach involves a group of experts participating in an anonymized process of two or more rounds (Roper et al. 2011). After each round a facilitator provides the results and an anonymized summary of all participants explanations (Beard et al. 2020). The falloff rate of participants over iterative survey rounds is the primary challenge when using this technique. The Delphi technique is powerful and versatile, being able to be used with slight modifications for large groups of experts in scenario building exercises (Amer et al. 2013).

There are two primary types of Delphi studies: quantitative Delphi studies for forecasting or the policy Delphi (Linstone and Turoff 1975). The quantitative Delphi for forecasting is the most common type of Delphi, and it is widely used in organizations (Rowe and Wright 2001). However, because experts are generally difficult to find for empirical investigations in academic studies on the effectiveness of the Delphi, little work exists validating its effectiveness. The policy Delphi is a significant departure from the quantitative Delphi and can be used in numerous ways for eliciting and aggregating qualitative expert opinion when group consensus is necessary among experts (Turoff 1970). Generally, the Delphi is considered one of the more powerful techniques which utilize expert opinion, and for a variety of applications and contexts the Delphi is commonly thought of as the method of choice for quantitative forecasting (Green et al. 2015).

The Delphi has been used for numerous technology forecasting applications. Some of these applications include the forecasting of the potential risks from future drugs (Møldrup et al. 2001) as well as a variety of applications related to energy sources and supply such as forecasting energy sector development (Czaplicka-Kolarz et al. 2009) and opportunities and obstacles for future small-scale renewable energy projects (Varho et al. 2016). The Delphi has also been used for forecasts related to other potential GPTs such as for forecasting nanotechnology impacts (Glenn 2006), as well as for hybrid studies that combine the Delphi with scenario analysis (Tseng et al. 2009).

3.2.3.3 Prediction Markets

Prediction markets are exchange traded markets intended for predicting the outcomes of events. Prediction markets employ a platform that allows people to make trades depending on their assessment of these outcomes (Beard et al. 2020). Prediction market contracts, which are binary options, are traded through the market. Through trading, the market price of a contract adjusts dynamically to account for participants' predictions and is used as an indicator of the probability of these events. This incentivizes participants to be as accurate as possible in order to receive the most gain while allowing for aggregation over an arbitrarily large market. For example, given a hypothetical election, consider a contract that will pay $1 if Candidate X wins. If the current market price for Candidate X contracts is at 47 cents, this can be interpreted that the market gives Candidate X a 47% chance of winning (Arrow et al. 2008). The free market tends to collect and aggregate predictive information well due to the strong economic incentives for better information.

There is strong evidence that prediction markets can produce forecasts of event outcomes with a lower rate of prediction error than conventional forecasting methods involving expert elicitation (Arrow et al. 2008). Prediction markets are also thought to be good options for practical use in improving business and political decision making (Hahn and Tetlock 2005, Tetlock 2017). They have been used by the US Department of Defense, Google, Microsoft, IBM and Intel[55] to employ expectations about organizational issues to inform decisions. Metaculus[56] is an online prediction market platform that uses a point system for forecasting a range of questions, some related to AI forecasting issues. One issue that poses a significant challenge to using prediction markets for AI forecasting is the development of near-term likely targets, or events, for using the market to predict (Dafoe 2018). Prediction markets are thought to have a wider range of practical applications for forecasting when compared with the Delphi technique, however, each method is best suited for particular uses and contexts (Green et al. 2015).

3.2.3.4 Superforecasting

Superforecasting is a recently developed technique that utilizes groups of forecasting experts, i.e., superforecasters, in combination with advanced aggregation techniques to generate forecasts. Superforecasting has been demonstrated to be more accurate than prediction markets (Tetlock and Gardner 2016), suggesting that it is the most accurate forecasting technique for certain applications. Superforecasting was developed by Phillip Tetlock as part of a competition for the US Institute for Advanced Research Projects Activity (IARPA) intended to improve the US intelligence communities' forecasting abilities (Schoemaker and Tetlock 2016). Tetlock had long been a proponent of forecasting competitions (Tetlock 2017), so, for the IARPA challenge, he created a competition that pitted thousands of amateur forecasters against career intelligence analysts. The results of the competition yielded three interesting observations: 1) talented amateurs[57] can outperform career specialists 2) specialized training can increase forecasting ability

[55] The Delphi technique may be used much more widely, but with experts and for proprietary forecasts. prediction markets are public and only used by organizations with substantial resources, so their use is better documented.

[56] www.metaculus.com

[57] Note that superforecasting techniques do not use domain experts.

and 3) carefully selected teams combined with novel aggregation algorithms can outperform individuals.

Superforecasting is not suitable for all forecasting problems. Particularly, it is ill-suited for predictions that are either entirely straightforward and well suited for econometric methods, or predictions that are seemingly impossible (Schoemaker and Tetlock 2016). Schoemaker and Tetlock use more specific language stating that they "find issues that are complex, poorly understood, and tough to quantify, such as … when the next game-changing technology will pop out of a garage in Silicon Valley" to be poor targets for superforecasting. They conclude that for such cases "The problems are just too hard to crack." Superforecasting is the most powerful forecasting method to have been evaluated, yet it still may not be able to make forecasts any more accurate than a coin toss for horizons over five years in the future (Tetlock and Gardner 2016). Thus, we believe it is significant that two of the world's foremost experts in political and organizational forecasting emphasize the analogue of AI forecasting to be effectively intractable without the inclusion of expert opinion. This is a significant motivation for the development of the new method and framework that are emphasized in this study.

However, there is a sweet spot; a class of forecasting problems for which superforecasting excels (Schoemaker and Tetlock 2016). These are forecasting problems in the range of months up to three years (Tetlock and Gardner 2016). Specific forecasting training, the use of superforecasting teams and the specialization of training and techniques to the unique needs of each organization can produce significant benefits on these types of problems. A custom, one-hour online calibration training module was administered during forecasting tournaments using superforecasters and other talented forecasters. This calibration training program was found to increase forecaster's performance by 10% or greater (Moore et al. 2017). Also, superforecasting requires scoring forecasts in order to rank and weight forecasters. The most common score for this purpose is the Brier score (Brier 1950).

We believe that this technique could be significant for AI forecasting, despite the inability to forecast effectively past a five-year window. Given an accurate mapping of the AI technical landscape (Dafoe 2018) in the sense we consider for this study, we could use such forecasting techniques to forecast intermediate scenarios, or technologies. This approach could also yield more powerful near-term forecasts into which extrapolated trend models or other expert elicitation timelines could be integrated. Much work remains to be done in this area, and whether or not the proposed technique for integrating forecasts is viable is a suitable question for future research. However, the use of superforecasting techniques would be beneficial in informing scenario mapping models with high quality timelines for at least the first layer of mappings.

3.2.4 Expert Selection

To this point we have only discussed the how of expert elicitation. The who, i.e., whose expertise is to be considered, is also a key consideration. This is not a problem for prediction markets, which does not require experts, or for superforecasting where experts are selected based purely on their previous performance and calibration not their domain expertise. However, for other forms of expert elicitation the selection of experts can be crucial.

The first step in identifying experts is to identify the range of perspectives that will be needed in the study (Roper et al. 2011). Researchers typically want to prioritize the most knowledgeable experts for vital perspectives first; less vital viewpoints can often substitute secondary sources for expert opinion. Researchers should also be cognizant of possible sources of experts' biases when selecting experts and analyzing their responses. Lipinsky and Loveridge (1982) suggest some desirable attributes for selecting experts for forecasting such as: a broad perspective relating to their knowledge of the innovation of interest, a cognitive agility for being able to extrapolate from their knowledge to satisfy future possibilities and uncertainties and a strong imagination.

There is also the question of how many experts one needs for a study. This is commonly dependent on many factors, including the type of the study, the technology of interest and the scope of the study. Sampling diverse populations can lead to many issues (Roper et al. 2011). However, when this is necessary documentation for the particular type of study commonly addresses issues of expert selection (Rea and Parker 2014).

3.2.5 Scenario Planning

Scenario planning is a technique that involves the development of a number of different scenarios for plausible possible futures. It is most widely thought of as a qualitative technique for the purposes of strategic planning in organizations (Roxburgh 2009). Proponents of this thinking often consider scenarios as an aid for thinking about the future, not for predicting it. However, a rich body of literature has developed over the years and quantitative and hybrid techniques have been shown to be practically useful (Amer et al. 2013). Here we attempt to outline some of the most widely used techniques and highlight their potential advantages for use in forecasting TAI.

The most prominent of qualitative methods, having received the most attention in the scenario planning literature, is the intuitive logics school of methods (Amer et al. 2013). It was first developed at the RAND Corporation in the 1960s and was popularized from its use by Royal Dutch Shell in the 1970s – it is sometimes referred to as the ‘Shell approach’ for this reason (Bradfield et al. 2005). This school of methods is founded on the assumption that business decisions rely on a complex web of relationships including economic, technological, political, social and resource related factors. Here, scenarios are hypothetical series of events that serve to focus attention on decision-points and causal processes. While such scenario planning techniques are very useful for business purposes, alternative scenario planning techniques can be used for much more than investigating blind spots in organizations' strategic plans (Chermack 2011).

The most common of quantitative methods are considered to be the probabilistic modified trends (PMT) school, which also originated at RAND Corporation in the 1960s (Gordon and Helmer 1964, Amer et al. 2013). This school incorporates two distinct methodologies: trend-impact analysis (TIA) and cross-impact analysis (CIA; Bradfield et al. 2005). TIA is a relatively simple concept which involves the modification of extrapolations from historical trends in four relatively simple steps. CIA attempts to measure changes in the probability of the occurrence of events which could cause deviations from extrapolated trends through cross impact calculations. The primary difference between the two techniques is the added layer of complexity introduced in CIA by the cross-impact analysis.

The two schools described above to illustrate qualitative and quantitative scenarios are by no means an exhaustive description of this dichotomy of scenario planning methods. Another way to think of qualitative and quantitative scenarios is as storylines and models. The former captures possible futures in words, narratives and stories while the latter captures possible futures in numbers and rules of systems' behaviors. Schoemaker (1995) suggests that the development of quantitative models is an auxiliary option for assisting in making decisions, whereas the development of scenarios is the purpose of the activity. Hybrid scenario techniques attempt to bridge the gap between methods which rely on storylines and models.

La Prospective is a school of hybrid scenario techniques that emerged in the 1950s in France for long term planning and to provide a guiding vision for policy makers and the nation (Amer et al. 2013). This school is unique in that it uses a more integrated approach through a blend of systems analysis tools and procedures, including morphological analysis and a number of computer-aided tools (Bradfield et al. 2005). Although it arose independently, this school can also be seen to combine, to a large extent, the intuitive logics and PMT methodologies. However, a full review of scenario planning literature is beyond the scope of this work, and we believe the simple descriptions used here to be sufficient for our purposes.

Traditional qualitative scenario planning techniques certainly have a role in assisting decision makers of organizations and other stakeholders involved in the development of AI and TAI/DTAI/RTAI. However, such traditional techniques can do little to map the paths to HLAI or other forms of RTAI due to the large space of plausible paths. Consequently, there are limits to what such methods can do to assist organizations and research groups in determining the best routes forward for their own research efforts. Traditional quantitative methods certainly have a role in some organizational decisions as well. They are normally sufficient for strategic decision making, but for understanding and informing design decisions of a complex system they fall short.

3.2.5.1 Scenario Mapping

Over the past two decades, the use of scenario planning techniques for mapping complex systems, complex environments and complex technologies has increased (Jetter and Kok 2014). In reviewing this literature we identified a class of scenario planning techniques that was well suited to the mapping task described earlier (i.e. mapping the AI technical landscape; Dafoe 2018). We call this class of methods scenario mapping techniques because of these mapping characteristics. In fact, each of the techniques includes map in its name. Moreover, these techniques all share the two significant properties: 1) they do not have a strict limit on the number of scenarios they can accommodate and 2) they represent the scenarios as networks with directed graphs. When represented as graphs, the nodes are scenarios that are connected by edges, which themselves are representative of causal relations between the scenarios. These properties of the techniques are starkly different from the vast majority of scenario planning techniques[58] (Bradfield et al. 2005, Amer et al. 2013).

This study focuses on three such techniques. The first is a relatively obscure method that has seen little practical application, yet it has been included because it has significant potential for mapping the AI technical landscape. The second originated as a way to represent social scientific knowledge through directed graphs and has since become a common method for scenario planning in multi-organizational contexts (Soetanto et al. 2011). The third extends the second by making those methods computable for quantitative forecasting, but also has practical uses in a large number of applications across various other domains (Jetter and Schweinfort 2011). Each of these techniques has potential for mapping the AI technical landscape and possibly quantifying forecasts. These techniques are described in more detail in the paragraphs below.

The first of these techniques that we consider is scenario network mapping (SNM) (List 2005). SNM is a qualitative technique that was developed in order to accommodate more scenarios than traditional methods (i.e. roughly an order of magnitude more: 30-50 scenarios are typical for SNM while 2-8 scenarios is typical for traditional methods). The technique is commonly more concerned with the network structure of the mapping than the actual scenarios generated. SNM uses multi-day workshops of roughly 20 participants for the scenario building process, which focuses on generating clusters of scenarios rather than independent scenarios (List 2007).

[58] Note that the scenario mapping techniques identified are not typically used for mapping technical components of emerging technologies. Rather, they are used for understanding the many moving parts of dynamic systems in the emergence of advanced technologies, e.g., economic factors, market forces, the rate of research progress in component technologies, etc. However, the development of HLAI is in a class unto itself (Gruetzemacher 2018) in that it is attempting to create an artificial mind. For these reasons, technology roadmapping fails. Despite their common use for more systemic planning problems, scenario mapping techniques are uniquely suited for mapping paths to HLAI or other forms of RTAI. Ideally, this is an intuitive notion. Technology roadmapping, while popular with more traditional technologies, failed in the previous attempt to use it for mapping HLAI because it does not capture the intricate network of different possible paths (Goertzel 2014b, Goertzel (2016).

Following the development of the mapping, the scenarios can be refined further using causal layered analysis, a powerful analysis technique for strategic planning and futures studies (Inayatullah 1998). SNM is a lesser known technique, but it has significant potential due in large part to its incorporation of the holonic principle. The holonic principle means that each scenario exists simultaneously as a whole, as a part of a larger system and as a system comprised of smaller parts. An extension of SNM has demonstrated excellent results for forecasting complex networks of scenarios involving interactions between innovation at the micro and macro levels (Gaziulusoy et al. 2013). Similar extensions of this technique may be suitable for mapping the AI technical landscape.

The second of these techniques is cognitive maps, also a qualitative technique, which was first proposed by Axelrod in the 1970s as a way to represent social scientific knowledge and stakeholder beliefs in the form of a directed graph (Amer et al. 2013, Axelrod 2015). However, Axelrod's use of the term was derivative, and its true origin was from cognitive psychology (Tolman 1948). (Interestingly, Lake et al. (2017) refer to "cognitive maps" in advocating model-based reinforcement learning for modeling the way this notion of cognitive maps is thought to be used by the brain for planning.) In the past decade they have come to be used for the purposes of scenario planning, but in this context are commonly referred to as causal maps (Goodier et al. 2010; Soetanto et al. 2011). They are known to be effective for developing scenarios in complex multi-organizational cases. Furthermore, they are promising for understanding complex systems and see great potential for use in mapping cognitive processes in artificial agents.

Cognitive maps have not received much attention in scenario planning literature when compared to fuzzy cognitive maps (FCMs), which are simply a computable extension of cognitive maps that incorporate fuzzy logic (Amer et al. 2013). First proposed by Kosko (1986), they are thought to be better for integrating expert, stakeholder and indigenous knowledge by enabling the development of scenarios that can aid in linking quantitative with qualitative storylines (Jetter and Schweinfort 2011). While widely used for scenario planning, they have a wide range of other applications across numerous disciplines that generally involve complex modelling and decision-making tasks, e.g., online privacy management, decision support, knowledge representation and robotics (Papageorgiou 2013). FCMs are weighted directed graphs with nodes that are fuzzy and representative of scenarios, or concepts, and with edges that represent causal relations. FCMs can be used to generate quantifiable forecasts, but their most significant feature may be their ability to integrate a wide variety of information types including subjective expert knowledge as well as technology, innovation and economic indicators. These properties can likely be extended to other scenario mapping techniques, or FCMs could possibly be adapted for mapping the HLAI technical landscape.

It has been suggested that expert interviews are a suitable technique for eliciting expert opinion to generate FCMS, and rough outlines and suggestions have been made for how to proceed when using interviews for developing scenarios and cognitive maps (Jetter and Schweinfort 2011). One suggestion is to split the process of capturing expert knowledge into separate activities. For example, a first step could utilize interviews to identify concepts and to clarify their meanings. A second step would then use interviews to clarify concept meanings. Finally, knowledge structures can be elicited through further interviews. The result of this process would be the creation of cognitive maps for individual experts, which would then need to be combined.

3.2.5.2 Elicitation of Expert Opinion for Scenarios

Virtually all scenario planning techniques use expert opinion in some way, and there are various ways in which expert opinion is elicited for scenario generation. These techniques include interviews, panels, workshops and the Delphi technique (Amer et al. 2013). Many times specific techniques rely directly on the methods for elicitation of expert opinion being employed. For example, the proprietary Interactive Cross-Impact Simulation (INTERAX) methodology relies on the generation of a large database of the use of an ongoing Delphi study with close to 500 experts to maintain and update a database of approximately 100 possible events and roughly 50 trend forecasts. As previously noted, the Delphi technique is a particularly powerful technique that was developed at the RAND Corporation in the 1950s, as part of the same effort that resulted in the first scenario planning (Bradfield et al. 2005). List suggests that four half-day workshops with 20 experts is optimal for creating SNMs based on his six case studies (List 2007). However, some techniques do not rely specifically on one method for elicitation of expert opinion. FCMs can be developed using expert panels, workshops or interviews. In the case of using interviews, where combining expert opinions is required, all experts' opinions can be treated equally or expert opinions can be weighted based on some assessment of confidence in an expert's judgement (Jetter and Kok 2014).

3.3 Effectiveness of Technology Forecasting

Here we have presented a wide range of methods and techniques for technology forecasting and AI forecasting. However, the question remains as to whether or not these techniques are viable practically. We explore this question below.

Trends in technological improvement such as Wright's law, Moore's law and other experience curves have been widely used to understand the trajectories for how technologies change over time and for forecasting future technological progress through extrapolation (Magee et al. 2016). However, such forecasts that rely solely on social indicators are often not indicative of technological progress or innovation and are not effective for forecasting technological progress (Roper et al. 2011). Rather than following a curve of exponential growth, technological capabilities more commonly follow S-shaped growth curves, or simply, S-curves. Despite this, significant work continues to consider technological progress equivalent to social indicators for the purposes of forecasts whose practical use is primarily for economic-based planning and decision making. Work toward these ends has demonstrated this to be very effective for these purposes (Nagy, Farmer et al. 2013), although questions have been raised about the effectiveness of experience curves in the context of information technologies (Nagy et al. 2011). More recently, efforts have been made to extend these techniques to develop distributional forecasts for technological progress that are effective for predicting future values of economic indicators tied to technologies (Farmer and Lafond 2016, Lafond et al. 2018).

Measuring the effectiveness of other technological forecasting methods can be more difficult still. Mullins conducted a large study of historical forecasts including a database of 2,279 forecasts extracted from 300 forecasting documents (Mullins 2012). Qualitative trend analysis, gaming and scenarios were all found to perform better than a random guess, but this could not be confirmed statistically. However, the attributes of methodology and timeframe were statistically significant, indicating that quantitative trend analysis produced the best performing forecasts and that short-term forecasts were more accurate. The study further found that most forecasts are pessimistic in that they tend to overestimate the date at which a predicted technology will arrive.

4 METHODOLOGY

The previous two chapters covered a broad range of methods for technological forecasting as well as previous studies on forecasting AI progress and other literature relevant to historical technology development and its impacts on society and the economy. In this chapter, we describe plans to demonstrate and evaluate six specific techniques that are relevant to this study; a proposed new method and framework that are also components of the methods are developed in the following chapter. Because of the use and comparison of different methods, this section will be broken down into seven sections on methods: an AI practitioner survey, tech mining, extrapolation, structured interviews, a scenario mapping overview, scenario network mapping and a Delphi study. In the final section we discuss the evaluation of these techniques. The following paragraph gives a brief overview of the elements of the study.

The study will explore seven different techniques for creating forecasts: an expert survey, the Delphi technique, tech mining, extrapolation, structured interviews, scenario network mapping and an alternative scenario mapping technique (proposed in the following Chapter). 1) The first of these has been used for AI forecasting, but this study will focus on using the technique for forecasting different forms of TAI through a novel questioning framework and through enhanced analysis. 2) Tech mining will be used to demonstrate its unique applications for forecasting purposes. 3) Extrapolation will also be demonstrated after a strong indicator has been identified for which this simple forecasting technique is actually useful. 4) Structured interviews will also be used in what is sometimes referred to as the mixed methods portion of the survey portion of the broader study. The next two techniques are both scenario mapping techniques. 5) This involves a novel workshopping process adapted from previous SNM efforts to meet the unique needs for mapping the paths to notions of AGI. 6) This involves another novel scenario mapping technique that is loosely based on FCMs utilizing Monte Carlo simulation. It is referred to as judgmental distillation mapping and is detailed in the following chapter. 7) The Delphi technique will be used to develop a research agenda for AI forecasting by using a Delphi variation (the process will inform the future use of this adapted Delphi for sourcing expert opinion regarding identifying the most salient forecasting targets). Excluding the survey and tech mining, these methods have not been used previously in the context of AI forecasting.

We develop a novel forecasting framework referred to as a holistic forecasting framework. This is also detailed in the next chapter. The framework is presented here to demonstrate the viability of the concept and is intended simply as one example of how such a holistic framework could be manifested. An alternate manifestation is proposed in Chapter 9, which builds on the results of the Delphi for contributing to an existing forecasting framework that meets the criteria of a holistic forecasting framework presented in the following chapter.

4.1 Expert Survey

An expert survey was conducted to forecast extreme levels of labor displacement from AI. This was part of a convergent parallel mixed methods design (Creswell 2014) that also included structured interviews using the same questions (discussed in Subsection 4.4). The survey was administered in the summer of 2018 at the International Conference on Machine Learning (ICML), the International Joint Conference on Artificial Intelligence (IJCAI) and the Joint Multi-Conference on Human-Level Artificial Intelligence (HLAI)[59].

[59] ICML and IJCAI were held consecutively in Stockholm, Sweden from July 10th-19th. HLAI was held from August 22nd-25th in Prague, Czech Republic. The HLAI conference was a joint conference between the Artificial General Intelligence Conference, the Biologically Inspired Cognitive Architectures (BICA) conference and the annual workshop on Neural-Symbolic Learning and Reasoning (NeSy).

The survey (Appendix A & B) was designed to generate forecasts for different forms of TAI. Consequently, it focused on five forecasts for different levels of extreme labor displacement that AI could lead to as well as other questions. In total, there were 13 subject matter questions in the survey. The first four questions (Q1-Q4) were calibration questions, intended to be used not for forecasting but for assessing respondents' forecasting skills when aggregating the forecasts. The next question (Q5) involved forecasting the portion of human tasks that would be automatable at present, in five years and in ten years. The following five questions (Q6-Q10) were the focus questions, which each asked for how long it would take before AI could complete a portion of human tasks at or above the level of a typical human. These questions asked for forecasts at three quantiles of 10%, 50% and 90% for extreme levels of labor displacement scenarios ranging from 50% to 99%. The last three questions all concerned computational resources. For the first two of these (Q11 and Q12), participants were shown a figure from Amodei and Hernandez (2018) depicting an extrapolated technology indicator and were asked to give a probability that this trend would continue for 5 years and for 10 years. The final question simply asked how much participants expected the rate of progress in AI research to increase if all researchers were given access to unlimited computational resources[60]. The complete survey instruments[61] are included in Appendix A and Appendix B for each different location, Stockholm and Prague, respectively.

The four calibration questions were selected as best as possible to be near-term probable events that could be unambiguously resolved within four years (Dafoe 2018). Ideally, targets would be selected that could be evaluated in a shorter time period in order to weight the forecasts, but this was not possible to do while requiring minimal domain specific expertise in forecasting;

[60] This was intended to explore the degree to which computational resources were perceived by researchers as a limiting factor in the rate of AI progress.

[61] Two separate instruments are included due to a minor modification in the 3rd calibration question and the use of an entirely different 4th calibration question between the two locations.

domain specific forecasting targets for short timelines would have been easier to generate. The targets that were selected for these questions were anticipated to be perceived as having a reasonable chance of both failure and success. Two of the questions concerned self-driving vehicles, while a third question concerned the academic productivity of DeepMind for the remainder of 2018 and the final question concerned an even nearer-term target that was unique for each of the conference locations. For each of these questions, participants were given a future scenario and a date, then asked to give a probability that the given scenario would occur.

Q5 asked for three point estimates[62] of the portion of human tasks that practitioners' anticipated being automated at present, in five years and in ten years. These forecasts are considered near-to mid-term forecasts (the questions are also referred to as fixed years[63] questions for the context of this study). These questions offer a means for gauging the overall results with respect to existing future of work studies that have made forecasts for near-to mid-term human task automation from AI.

For the five extreme labor displacement scenarios, rather than specifically discuss notions of AGI, we refer to narrow systems and broadly capable systems. This was considered more appropriate for several reasons. First, because TAI/DTAI/RTAI may not be manifest in the ways we currently envision it and these terms are intentionally generic. Second, HLAI, AGI and similar terms are buzzwords in the AI community for which many researchers have strong opinions that may have the effect of biasing responses or increasing nonresponse bias. Thus, narrow systems are defined as systems which specialize in one task or a small domain of tasks and broadly capable systems are defined as systems that can accomplish a large range of tasks through a single

[62] A simple probability or percentage.

[63] These are called fixed years questions because they ask for a point estimate at a fixed point in time. Alternately, the questions on extreme labor displacement give a fixed probability and ask for a year.

algorithm[64] or agent. We considered these two system types for three levels of human task competition, i.e., 50%, 90% and 99%. Specifically we considered five different extreme labor displacement scenarios: 1) narrow systems capable of completing 50% of human tasks, 2) narrow systems capable of completing 90% of human tasks, 3) a broadly capable system able to complete 90% of human tasks, 4) narrow systems capable of completing 99% of human tasks and 5) a broadly capable system able to complete 99% of human tasks. Human tasks were defined as all unique tasks that humans are currently paid to do. Human tasks are differentiated from jobs in that jobs are comprised of human tasks, as is consistent with the work discussed in the literature review (e.g. Autor et al. 2003).

The final resource related questions each were intended to evaluate the suitability of compute as an indicator of AI progress, as has been suggested by some (Brundage 2016). Amodei and Hernandez (2018) plotted a technology indicator for AI in 2018 which depicted the amount of computational resources used to achieve major AI milestones. This plot had generated a lot of interest in the AI community and was recognizable to many who were shown the plot during structured interviews. After being shown the plot, respondents were asked for a probability that this trend would continue for five years (Q11) and ten years (Q12)[65]. Following this, participants were asked how much they believed progress in AI research would increase if researchers were given access to unlimited computational resources (Q13). Participants were limited to one of seven responses for this question: it would slow down, 0%, 25%, 50%, 100%, 200% and 400%. This

[64] i.e. through the optimization of a single utility function.

[65] These questions are also fixed years questions, but in the context of the study they are referred to as compute questions.

question was included as an attempt to try to gauge how significant a bottleneck practitioners felt computational resources to be with respect to current progress[66].

Participants were selected at random from the three conferences and solicited for participation in the survey. Agreeable participants' email addresses were collected and a link to the survey was immediately sent. Two follow up emails were sent to participants who had not yet completed the survey: the first at the end of the conference and the second three days later. Participants were asked to self-report their employer and their job title along with their names. This basic information was used to identify and remove non-practitioners[67] from the sample. Participants who completed the survey received a $15 Visa gift card.

The objective of this study as part of the broader study is to generate a forecast for levels of TAI (as well as to generate information that can be used to inform the technique and framework described in the following chapter). Because TAI is possible with existing technologies, the goal is to forecast either DTAI[68] or RTAI. Because this technique is utilizing expert judgement, and expert forecasts are unreliable for long time horizons (Tetlock and Gardner 2016), the goal of this demonstration will be primarily to generate forecasts for DTAI. Thus, while longer time horizon forecasts will be generated, the focus of the results and analysis will be the near- to mid-term DTAI scenarios (i.e., extreme labor-displacing AI scenarios of 90%) that have not been explored in any prior forecasts or work on the topic of AI progress or the future of work.

[66] Many in academia or poorly funded research labs have complained about not being able to conduct research at the level of major tech labs like Google or Microsoft due to the extreme amounts of computational resources such groups are able to dedicate to AI research. The significance of computational resources on research progress was made clear by the plot from Amodei and Hernandez (2018), so this question was a simple extension of the theme from the previous two questions. Notably, Brown et al. (2020) demonstrated the power of computational resources and model scaling with the few-shot learning of the GPT-3 model. Sutton (2019) has also written on this topic, and it was discussed in his interview prior to the blog post.

[67] Non-practitioners are considered as those participants whose occupations do not fall into any of the AI practitioner categories, e.g., human resources, sales, journalism, etc.

[68] Which is also potentially possible with existing technologies.

4.2 Tech Mining

Tech mining techniques are demonstrated here because they are a very widely used technique in technological forecasting (Roper et al. 2011). However, because the technique is so labor intensive – i.e., it requires a large number of hours scraping, aggregating and cleaning data in addition to the analysis – this demonstration does not seek to generate a forecast[69]. Rather, it will be used to conduct an exploratory analysis, to evaluate the impact of data quality on the results and to inform the scenario mapping techniques. To be certain, this is its designated purpose in the holistic forecasting framework described in Chapter 5, and the exercise here to demonstrate this is sufficient for the purposes of this study.

As noted, tech mining techniques will be used here for a simple, exploratory study. Ideally the study would like to identify useful indicators and other topics of interest (Porter and Cunningham 2004) that may be effective in the framework presented in Chapter 5. For this demonstration, and to evaluate the impact of data quality, three datasets will be used. The first will be collected in the form of abstracts from the Web of Science database using keywords to search for document titles containing "AI" or "artificial intelligence" in any publication included in the database. A second dataset of machine learning abstracts will also be collected by searching Web of Science for documents titles containing a set of terms selected to ensure inclusion of topics from each of the five different schools of machine learning (Domingos 2015): 'neural network,' 'deep learning,' 'machine learning,' 'support vector machine,' 'inverse deduction,' 'backpropagation,' 'genetic programming,' 'probabilistic inference' or 'kernel machine.' Each of the datasets obtained

[69] The Center for Security and Emerging Technology, a think tank affiliated with the University of Georgetown, employs a data science team of at least ten personnel who all focus on this task in an effort to construct a usable dataset for accurate forecasting of AI progress. Thus, conducting any demonstration of this technique which can be used to generate even nominal forecasts is beyond the scope of this study. However, this anecdote should suffice to underscore the value that many who are heavily invested in technology forecasting place on this class of techniques.

by querying Web of Science will include data dating to as early as 1987 (if possible). The third dataset to be utilized includes all full text papers published in the proceedings of the Neural Information Processing Systems (NIPS) Conference from its inception in 1986 until 2016.

The nominal objective of this exercise is to identify trends or hidden patterns from an exploratory search of the datasets described which can be used for informing the other methods demonstrated in this study. Because actual forecasts are beyond the scope of this demonstration, and not relevant to the application of this technique in the framework presented in the following chapter, the underlying objective of this portion of the study is to try to glean insight into some unique challenges that may be posed using this class of techniques for generating actual AI forecasts. This can be practically useful to other researchers because, at the time the work was conducted, there had been no published work conducted on this topic which actually attempted to create forecasts or valuable insights about AI progress using mined data.

4.3 Indicator Extrapolation

This demonstration was included in the study because it is historically the most widely used forecasting technique and also commonly one of the most successful forecasting techniques (Roper et al. 2011). For AI forecasting, it is the one technique that has been used successfully in the past (i.e., when Russell and Norvig (1995) used it to project the year when AI systems would obtain an ELO score equivalent to the world's best chess player). It is also a technique that received a lot of attention in 2018 with the release of the Amodei and Hernandez (2018) plot depicting the correlation of computational resources with AI milestones.

The objective of this demonstration was to identify a strong indicator that could be used to forecast progress in AI toward some level of TAI. As existing AI technologies constitute TAI, the goal is thus to forecast DTAI or RTAI. This is an ambitious goal, and despite the lengthy definitions in Chapter 2, both DTAI and RTAI are poorly defined for the purpose of being practical forecasting targets (Dafoe 2018). Consequently, the goal for this forecasting method is to forecast RTAI since trend extrapolation is well suited for forecasting on longer time scales (e.g. Moore's law).

4.4 Structured Interviews

Interviewing as a qualitative research technique attempts to understand the world from the subject's point of view simply through the means of conversation (Kvale and Brinkmann 2009). However, in contrast to many quantitative methods, no standard rules or procedures exist for conducting an interview or an entire interview investigation. Some idealized steps have been proposed, but they are not commonly applicable in practice. One general rule does hold, though; the better prepared the interviewer the higher the quality of the knowledge produced by the interview.

The structured interviews conducted for this study are part of a converged parallel mixed methods approach (Creswell 2014) that is intended to create forecasts for TAI as well as to explore unique nuances of expert elicitation in the context of AI forecasting. Structured interviews are a common survey technique frequently used for eliciting expert opinion in technological forecasting applications (Baker et al. 2009, Bistline 2014). The structured interviews used in this study, and as part of the mixed methods approach, will utilize the same questions as the survey. However, the structured interviews also contained three further questions listed below:

1. What do you think are the most significant AI milestones in the past two years?
2. What milestones do you anticipate as being indicators of true progress in AI over the next five years?

3. You have just made forecasts for AI systems capable of completing 90% and 99% of human tasks. What kinds of systems or milestones do you anticipate preceding the development of such forms of advanced AI?

The interviews were conducted at the conferences where participants for the survey were solicited, as well as at the 2018 Neural Information Processing Systems (NeurIPS) conference held in Montreal, Quebec. Experts were selected for invitations to participate in interviews if they were listed as invited speakers for the current conference or a workshop associated with the conference of attendance. Research scientists from research labs working on notions of AGI were also extended invitations if they were known to be attending the conference. Given these criteria, hundreds of invitations were sent in total. Invitations were sent slowly, beginning a few days before each conference and continuing to the middle of the conference, at which point the list of potential candidates was exhausted. No follow-up emails were sent to invitees who did not respond[70]. Participants for any of the interviews in this study received a $40 Visa gift card upon completion.

Structured interviews were used in this study to satisfy two primary objectives. First, the interviews can be used to inform the preliminary stages of the holistic forecasting framework that is described in the following chapter. This is a necessary component of the framework and thus is necessary for demonstrating its viability. Second, the interviews can be used in the mixed methods framework to assess the value of expert opinion in studies related to forecasting AI progress or TAI. Of particular interest is the impact of expert selection on the value of the data collected – e.g., are biases present in researchers new to the field, veteran academics, employees at AGI research

[70] Invitees were prioritized based on research foci and perceived recognition in the AI research community. Some of the most respected experts were not invited for structured interviews so that they may be contacted to participate in the later stages of this study or future studies.

labs, etc.?; is research focus associated with certain biases?;does face-to-face interaction impact data quality?

4.5 Scenario Mapping

Section 3.2.5.1 describes scenario planning techniques with qualities we consider to be characteristic of scenario mapping techniques. We further seek to use or extend these techniques to address some of the unique challenges of AI forecasting and forecasting TAI/DTAI/RTAI. As their name implies, these techniques could be useful in developing the mapping of the technical landscape proposed by Dafoe (Dafoe 2018). Similarly, these techniques could be used to map the possible paths toward notions of AGI, thus yielding valuable information about different research avenues that may be relevant to manifestations of RTAI or DTAI which are not easily classified as either some notion of AGI or a true version of CAIS[71]. Prior to this study, no such mappings have been created and no work has been conducted to develop such a mapping systematically.

This study explores two different scenario mapping techniques in order to better understand their viability and effectiveness for the purpose of AI forecasting. Moreover, they have the potential to complement each other when used in a manner consistent with a holistic forecasting framework. For mapping the paths to notions of AGI SNM will be used. This will require the development of a novel adaptation of the SNM workshopping technique to map the technology's paths directly, as opposed to using SNM for anticipating and mapping different social consequences of the development of technology and the pursuit of different policy trajectories. To develop a mapping of the AI technical landscape (that can also be used for quantitative forecasting) FCM will be used as a starting point for developing a novel technique better suited for AI

[71] A true version of CAIS is interpreted here to be AI systems that are truly comprehensive in nature, and which can accomplish all human tasks at a superhuman level. This could imply that the development of a broadly capable system using these technologies could be built but is not built due to the lack of incentives for creating such a system. This also implies an alternative to superintelligence (Drexler 2018).

forecasting and forecasting TAI. In this chapter, we focus on details of the SNM demonstration that is to be conducted for this study. In the following chapter we will consider FCM and develop a new method that we refer to as judgmental distillation mapping.

4.6 Scenario Network Mapping

SNM is a technique that was developed 15 years ago as a doctoral dissertation to address weaknesses of traditional scenario planning techniques (List 2005), as was discussed in the literature review. Gaziulusoy extended the technique for her doctoral dissertation five years later (Gaziulusoy 2010). We will use the technique here for creating scenario network maps of paths to notions of AGI by adapting from the List's original method and drawing from different alterations proposed by Gaziulusoy. To our knowledge, and despite suggestions for its viability in other applications, this technique has not been used for any purpose other than that for which it was originally demonstrated (i.e., mapping socio-political scenarios that could result from a technology and its development). However, List suggests that, with adequate effort, the technique can be extended to be used for exploring different technical elements of the development of a technology itself[72], as well as for modeling impacts of the technology development on society (List 2006). Thus, the effort here to develop the technique for exploring technical paths of technology development marks the first effort to extend the proposed workshopping techniques for scenario mapping into this new type of application.

SNM utilizes workshops for the development of scenario maps. The workshops consist of between six and twenty participants. For this study, three workshops with between six and twelve participants were conducted to create a scenario network map and to iterate on the new

[72] As mentioned in the previous Chapter, SNM is more commonly used to explore different elements of circumstances involved in the development of the technology, in the same context that most scenario planning techniques would be used.

workshopping process in order to improve it[73] so that it can be used with high-level AI scientists and researchers. Participants of all workshops were offered a $40 Visa gift card upon completion. The workshops consisted of four two-hour sessions on different days, two half-day workshops (i.e., a morning and afternoon), or four half-day workshops on consecutive days as permitted by the scheduling of the parent event. The details for planning each of the three workshops are discussed in the following paragraphs.

The first workshop was held in Spain with AI safety researchers at AI Safety Camp III in April of 2019. This camp was a nine-day event held for selected international participants to work collaboratively on teams for AI safety projects at a quiet countryside inn near Madrid. All participants had substantial knowledge of AGI and related technologies, but not all could be considered experts. Six participants began the workshop, but only five completed both days. The participants all had a technical background, but their backgrounds were representative of a large range of different types of technical expertise (from quantum computing to full-time AGI researcher). However, for the purpose of demonstrating and developing the technique, the participants were satisfactory. The resulting map, as with all of the forecasts in this study, is not intended to be acted upon, but rather as a demonstration and as a component for informing a meta forecasting analysis.

The plan for this workshop was developed primarily a priori. The plans for the remaining workshops were developed more extensively on site based on the circumstances and constraints of the event and venue, using the action research cycles approach advocated by List (2006). These experiences will be described in the results. In this section we will only describe the a priori plan for this first workshop. For this workshop, the schedule was set to be four two-hour sessions; two

[73] This iterative process involved action research (Davison et al. 2004), and followed in List's (2005) use of action research cycles to develop the original version of the workshop.

on the first day, with a 30-minute break between sessions, and two more two-hour sessions after a day of rest (List (2005) suggests rest days if possible). Each of these sessions was derived from each of the four different workshop formats that are included in List's (2006) user manual for SNM. Despite the a priori plan shown below, significant fluidity with respect to the workshopping process was required even for this first workshop, as consistent with List's action research cycles approach to development. Because the workshop outline below is simply a template for the first SNM test, we do not elaborate on the details of each step. In practice it quickly became a rough guide, and more practical approaches relevant to the context were required. More detail will be included regarding the particular steps utilized in the discussion of the three SNM workshops in the results.

- Workshop 1: Influences from past and present
    - Introduction
    - Unfinished business
    - Prouds and sorries
    - Scenarios of the recent past
    - Stakeholder map
    - Leaf of goals
- Workshop 2: Generating possibilities
    - Futures wheel
    - Defining paths
    - Backcasting
    - Midcasting
- Workshop 3: Mapping paths to the future
    - Introduction and review
    - Grouping the event trees
    - Linking the event trees
    - Reviewing and digitizing the scenario map
- Workshop 4: Revealing the underlying layers
    - From event trees to scenarios
    - Finding the influences
    - Grouping the stakeholders
    - Finding the visions
    - Finding the worldviews
    - Review

The second workshop was held in Germany in conjunction with an AI-focused retreat in July of 2019. No further details are given as the event was small and we wish to ensure participant confidentiality. This workshop was part of a six-day event held for discussion of different topics related to the future impacts of AI. The majority of participants in this developmental workshop were also not experts on AI or AGI, rather, their expertise was varied, but the most common trait was their advanced degrees related to neuroscience. So, while representing a diversity of views, there was also a possibility of sample bias. Ten participants took part in the workshops which involved two half-day workshops held on a single day with a break for lunch. An effort to focus more on safety elements of AGI development was attempted in this workshop, however, this was complicated due to various reasons to be discussed in the results and discussion chapters. More details regarding the workshop itself are included in these chapters also.

The third workshop was held in Moscow, Russia during November of 2019 at a co-working space roughly a half mile from Red Square. Danila Medvedev provided support organizing this workshop with futures experts from academia and industry in the Moscow area. The majority of these experts were non-technical and were representative of the futurist community in Moscow. A total of twelve participants took part in at least one of the four workshops. The average attendance was between seven and eight participants. Because the workshop was an independent event held on half days during a work week, it is understandable that many participants had to miss at least one session.

The host of this event, Danila Medvedev, is a leader in the transhumanist community in Moscow and is lead developer of a software platform, NeyroKod, which is designed for mapping future scenarios. Thus, this novel software platform was well-suited for the purpose of using SNM to map the paths to various notions of AGI[74]. Because this workshop was organized independently of any other event, this enabled us to conduct a full, four-day workshop involving four half-day mini workshops (as intended for the technique originally proposed by List).

The result of the workshop will be a scenario network map for the development of notions of AGI. The map will resemble Figure 4, which depicts the structure of a generic scenario network map in the context (i.e., an approximation of what we expect to see from the results of each workshop). The map will indicate possible future paths to differing notions of AGI in a manner that has not been presented before. This map will differ substantially from the mapping of the HLAI technical landscape presented by Adams et al. (2012) or the roadmap of Gil et al. (2019), because of the different class of technique used. We see substantial value being derived not only from developing these maps and gleaning insight into the process, but also in the intermediate technologies that experts reach consensus on for being critical to the development of advanced AI. We anticipate that such technologies may lead directly to DTAI, or possibly even what could be considered RTAI, without having to reach the objective of some notion of AGI.

[74] The software is not publicly available, but it is much better suited for the purposes of this study than any alternatives. This workshop opportunity was possible because Danila reached out after reading the AGI-19 conference paper and arXiv preprint on SNM (Gruetzemacher and Paradice 2019a, Gruetzemacher and Paradice 2019b).

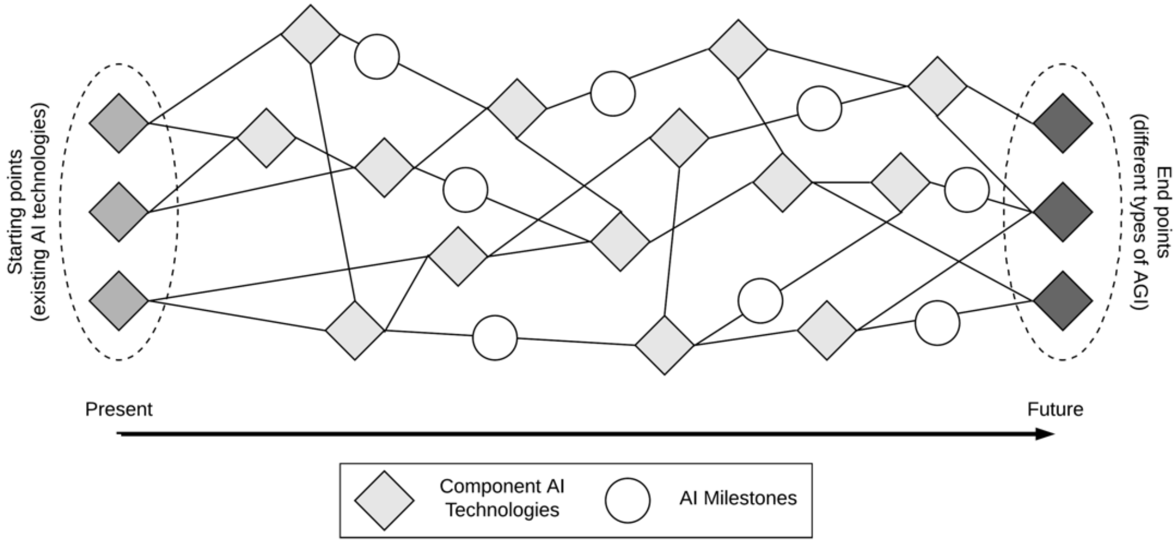


Figure 4: A generic scenario network map illustrated in the context of mapping the paths to AGI including the component AI technologies as well as an AI milestone layer for monitoring/assessing progress.

4.7 Delphi Technique

The Delphi technique was utilized also, however, it was not for forecasting directly. Rather, an adapted Delphi study was conducted in order to elicit the opinion of experts in the nascent and varied research area of AI forecasting. The output from this is a research agenda as well as rich data for inclusion in the future work chapter of this dissertation. Demonstration of this technique for this purpose will 1) be the first use of the Delphi in the context of AI forecasting and 2) will demonstrate its ability to successfully (or not) aggregate disparate and sometimes diverging views on this very broad topic for informing AI forecasters.

This Delphi process involves a single Delphi study that is inspired by non-quantitative Delphi processes like the policy Delphi (Turoff 1970) as well as by previous cases where the technique has also been used to generate a research agenda (Kellum et al. 2008, Dahmen et al. 2013). This particular Delphi process includes a first round that utilizes a questionnaire followed by an anonymous discussion stage conducted through Google documents. This round generated a summary of the results as well as two lists: one list of the most important questions for future research to consider and a second list including the most important methods to focus on for forecasting research. The second round involves the scoring of the items comprising each of the lists so that they may be ranked in the resulting report of the research agenda. This is similar to previous work which used the Delphi for research agendas in different domains (Dimmit et al 2005, Gordon and Barry 2006). This process lends itself well to other attempts to use expert elicitation for identifying the most salient questions of interest and ranking them (e.g., for identifying the most important forecasting targets relevant to a topic such as forecasting TAI) and will have implications on future efforts toward different ends.

Careful effort was made in the design of the process for the specific needs of this study, as well as in the development of the questions. Ultimately, only four questions were chosen for inclusion in order to minimize the potential for fatigue and nonresponse by experts. The Delphi questionnaire, including these four questions, is shown in Table 3. The first question is intended to explore experts' opinions on the tractability of forecasting AI progress as well as to determine whether experts believe that there is a clear path forward for forecasting AI progress. This is related directly to this study in that this study explores alternative techniques for forecasting AI progress – particularly forecasting TAI – because of shortcomings of existing techniques (Brynjolfsson and Mitchell 2017). The second question asks for the most important research questions and the third question asks for the most important research methods. These questions were necessary for collecting the opinions necessary for creating an objective research agenda for future researchers. The final question asks participants to identify ignored topics in order to effectively elicit opinions on other questions or methods that may have been missed due to the framing of questions two and three.

Table 3: Delphi Questionnaire

#	Questions
1.	Do you feel that forecasting AI progress is, or could be, a well-defined research topic? Why?
2.	What questions should researchers who work on forecasting AI progress prioritize?
3.	What methods or techniques should researchers use/prioritize to answer these questions?
4.	Are there any topics relevant to forecasting AI progress that you feel are important but neglected? Why?

There are only a small number of AI forecasting experts in the world, yet there is a broad range of diverse opinions among them. In order to represent the entire range of different opinions as many qualified experts as possible were invited to participate. Two criteria were applied when determining whether experts are fit for participation: the expert has written or published a rigorous or detailed blog post or academic article on a topic related to AI forecasting or evaluating AI progress, or, the expert is working full-time on practical efforts to forecasting AI progress. 31 experts met one of these criteria and were asked to participate in the study. More details regarding the experts who were invited and who participated can be found in Chapter 9, where the results of this study are reported in detail. Efforts were made to also include persons from other relevant disciplines such as economics or foresight who could reasonably be considered to meet these criteria (Rowe and Wright 2001).

The objective for the demonstration of this technique is actually three-fold. First, the objective is to demonstrate that the technique is valid and practically useful for topics such as AI forecasting where there is less shared understanding than for many topics that Delphi studies are applied to. Second, the goal is to create an objective research agenda to help guide the field (Gruezemacher et al. 2020); there is no document currently existing which can be used as a reference for people who are new to AI strategy or AI safety research, and who thus have a vested interest in AI forecasting. Such a document could be useful for this purpose as well as for aligning the study of AI forecasting more closely with academic research rather than with groups such as AI Impacts. Finally, the use of the Delphi in this form, i.e., the adapted Delphi process used, has implications for applications in alternative versions of a holistic forecasting framework discussed further in the next section. Particularly, the experience gained from this demonstration will be used to propose a modification for incorporating expert judgement with full-inference-cycle tournaments (Tetlock 2017) that constitutes another plausible holistic forecasting framework.

4.8 Method of Analysis

Table 4 depicts the techniques to be used in the study as well as their expected primary outputs[75]. The expert survey will generate probabilistic forecasts for five levels of extreme labor displacement (we focus on the implications of these forecasts on DTAI). The tech mining will generate an exploratory analysis of AI progress over the past 30 years while the Delphi will generate insights into the applications within an alternate holistic framework as well as an objective research agenda for guiding future research on the topic of forecasting AI progress. The extrapolation demonstration is expected to produce a direct forecast for RTAI (which, in light of the literature discussed in Chapter 2, also offers insight into DTAI). The structured interviews will provide both content for developing an initial mapping required for iteration in the judgmental distillation mapping process and probabilistic forecasts that enable a qualitative comparison with the survey results, as well as a thorough analysis concerning possible differences in biases using digital communication and face-to-face communication. The SNM workshops will yield scenario network maps of the paths to AGI as well as for the intermediate technologies and milestones on the way. The new scenario mapping technique will result in a scenario map generated from combining the outputs from the techniques demonstrated here that is further refined through rounds

[75] Institutional Review Board approval was obtained as appropriate for each technique demonstrated. Details regarding this can be found in Appendix H.

of iterative expert interviews. This map can be demonstrated to generate forecasts, although the near-term forecasts used to generate mid-to long-term forecasts will just be for illustrative purposes only. Completing the full holistic forecasting framework proposed would require more resources than were possible in this dissertation and is thus beyond the scope of this study. This is discussed as a limitation of the proposed methods because it suggests that only large, well-funded organizations are able to conduct high-fidelity forecasts via a holistic approach.

Table 4: The Components and Expected Results of the Study

Technique	Participants	Output
Survey of AI Practitioners	Over 100	Five forecasts for 50%, 90% and 99% levels of extreme labor displacement from AI, forecasts for near-to mid-term labor displacement and assorted compute forecasts
Delphi Study	6 to 20	A research agenda for AI forecasting and in informed process for incorporating the Delphi with alternate holistic forecasting frameworks
Tech Mining	n/a	Trends or other forecasting related information to use for informing decision makers and other forecasting efforts
Extrapolating Indicators	n/a	Efforts will be made to identify a strong technological or social indicator and extrapolate a forecast from it
Structured Interviews	<20	Five forecasts for 50%, 90% and 99% levels of extreme labor displacement from AI, near-to mid-term forecasts for labor displacement and assorted miscellaneous questions
Scenario Network Mapping	groups of 6 to 12	A novel SNM workshopping process specifically for mapping the paths to notions of AGI will be used to generate an SNM and facilitator notes for the workshop
New Scenario Mapping Technique	25-50	A new scenario mapping technique will be proposed and demonstrated (based on FCMs and which involves expert elicitation) by creating a forecast

Here we reiterate that none of the results from this study are intended to be taken as valid forecasts. Each technique performed was done so for demonstration purposes because all of the techniques are in some way different from any previous work conducted in the context of AI forecasting. However, this is not to say that they may not lead to valuable insights. Moreover, the results of the Delphi portion of the study are an exception, and the analyzed product is intended for dissemination and action. While none of the forecasts independently are intended to be considered as valid forecasts, this study would be somewhat incomplete without reporting, by some means, an overview of the foresight that was generated in the processes comprising it. Consequently, Chapter 11 is dedicated to a rudimentary discussion of the insights about the future that have arisen as a result of the many interviews, workshops, conferences and other interactions that the author participated in, as well as the results and the analysis conducted. However, this is not intended to be a rigorous examination.

A timeline for the completion of the different components of the study is shown in Figure 5. This is included to give readers an idea of the timeline for the various components of the study.

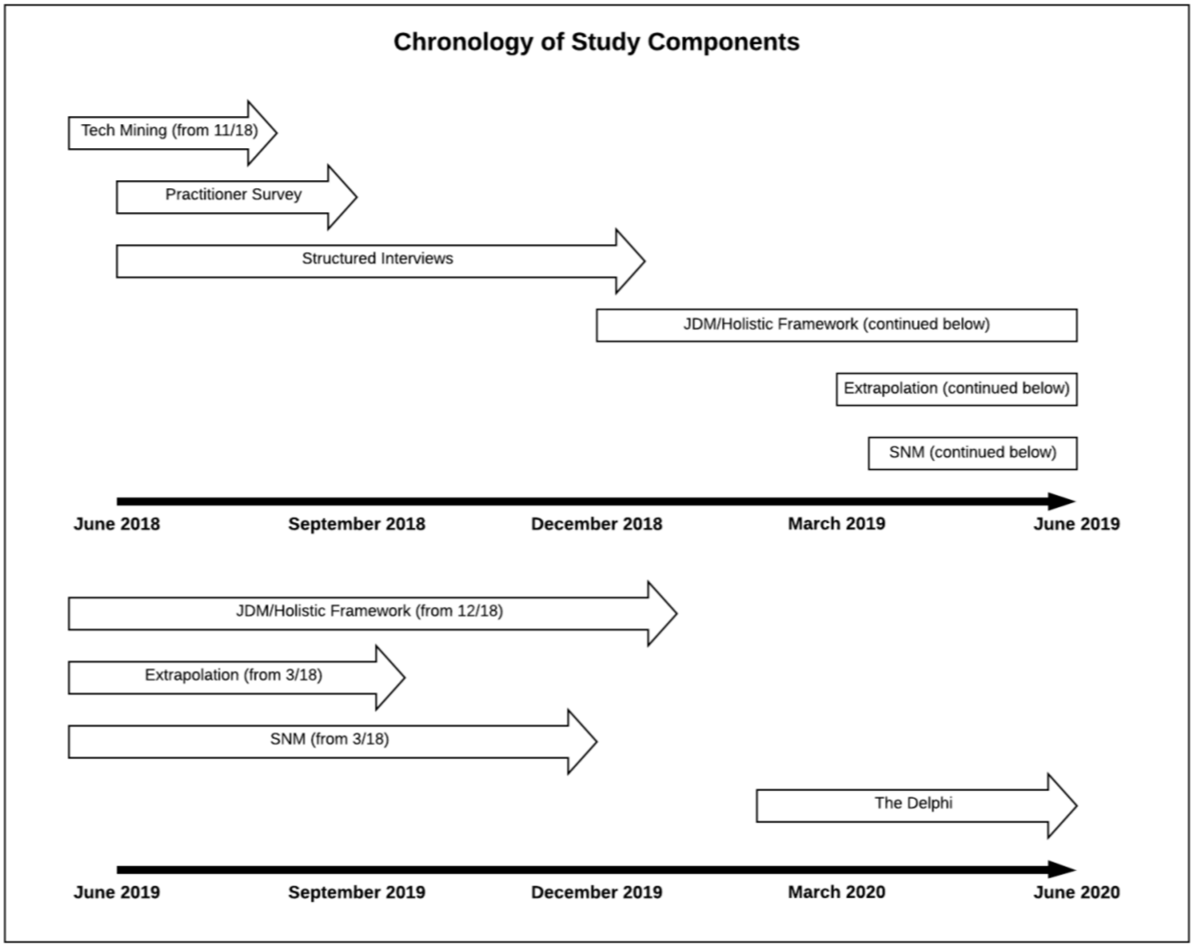


Figure 5: The chronology of the different components of the study.

4.9 Deviations from Original Proposal

Substantial changes were made to the plans for how the different methods were demonstrated following the dissertation proposal defense. The changes were necessary because of major insights gained during the progression of the study. The most significant of these was the development of the holistic forecasting framework and the judgmental distillation mapping technique; these were such a significant development that they constituted adding a new, separate chapter. The objective of the Delphi technique demonstration was also altered because it became apparent that more value could be derived in proposing an alternate holistic forecasting framework, because the bigger picture is that there needs to be a combination of expert judgement, scenario analysis techniques and other forecasting techniques in a holistic manner in order to generate effective forecasts for such a complex and challenging domain as TAI/DTAI/RTAI. Thus, a choice was made to explore the use of the adapted Delphi for obtaining consensus and ranking of salient questions from experts – this has significant value when combined with an existing, yet limited, holistic framework (Tetlock 2017). Moreover, using the Delphi to forecast the extreme labor displacement levels – as originally proposed – would have not been an optimal use of experts' time, and due to the limited availability of experts, there is a risk of nonparticipation in the future if they feel their time was poorly used. So, this could have had negative long-term effects on AI forecasting efforts. Specifically, granular, more accurate forecasts for the extreme levels of labor displacement from AI would not have been useful, because, among other reasons, improved accuracy could not be evidenced. These experts' opinions could be used more effectively to identify short term forecasting targets, or as part of the judgmental distillation process described in the following chapter. Other modifications include the addition of tech mining and extrapolation, which serve to demonstrate the two most common technology forecasting techniques as well as the latter representing the one known successful AI forecast. While these were not in the original proposal, they significantly increase the value of the study.

5 A HOLISTIC FRAMEWORK

Technology forecasting is a formidable task on its own. Forecasting TAI/DTAI/RTAI, however, is even more challenging because the techniques that are most commonly thought to be successful for forecasting other technologies are not necessarily the techniques that are best suited for forecasting TAI. The methods described in the previous chapters have focused on expert opinion but have also included more traditional methods including the use of extrapolation or tech mining, as well as more qualitative type methods such as scenario analysis. Generally, each of these techniques can be applied to a wide variety of topics related to technology forecasting, although they all are better suited for some types of forecasts than for other types. Extrapolation is a very powerful technique and has been famously applied very successfully to forecasting in the semiconductor industry or for forecasting superhuman level performance in chess playing AI systems, as mentioned previously.

Because different methods are better suited for different types of projects, one can logically deduce that, because forecasting TAI/DTAI/RTAI is such a complex and nuanced problem, multiple techniques are likely necessary for these purposes. However, aside from FCMs, there is little previous work on combining judgmental and statistical forecasting techniques, and FCMs have some severe limitations and do not appear to be optimal for hybrid or quantitative forecasting purposes. Consequently, a new method for combining forecasts is desirable.

However, simply a new method may not satisfy the severe challenges necessary for forecasting TAI/DTAI/RTAI. As Brynjolffson and Mitchell suggested, "a new framework is needed" (Brynjolffson and Mitchel 2017). Consequently, a new framework is proposed here.

Based on the logic previously outlined and the research in this study a new framework that involves the use of multiple types of significantly different methods has been developed: a holistic forecasting framework.

In this chapter we describe a holistic forecasting framework that includes a new method. This new method, called judgmental distillation mapping (JDM), is also described and a proof-of-concept is demonstrated. Briefly, JDM is akin to a combination of the Delphi technique and FCMs. The holistic framework utilizes JDM as well as all of the other techniques explored and evaluated in this study.

Ultimately, the specific holistic framework proposed in this study is but one possible form that a holistic forecasting framework could take. Only one other example was identified which also meets the criteria that are described for a holistic framework (see section 5.2). This example is still just a concept, like the holistic framework proposed here, and has only been described in an unfunded 2017 grant proposal (Tetlock 2017). At the end of the chapter, this example is explored briefly and the possibility of new and alternative frameworks that meet the holistic criteria proposed here are also discussed.

5.1 Judgmental Distillation Mapping

FCM is an extension from cognitive mapping that is widely used for a broad range of applications. The technique has become popular for scenario planning purposes over the past decade but has not been widely used for the purpose of technology forecasting. While previous work has suggested that it is suitable for quantitative forecasting purposes as well as qualitative forecasting and scenario analysis (Amer et al. 2013), there is little evidence to suggest that it is useful for generating quantitative forecasts that are practically useful (Jetter and Schweinfort 2011). The notion that experts' judgements can be combined with a rigorous methodology to generate probabilistic forecasts is an appealing one, even if FCMs are not able to do it in the manner necessary for informing decision makers with actionable probabilistic forecasts. Ultimately, there is to be a gap in the existing literature concerning a method for such cases. Because informing decision makers with actionable probabilistic forecasts is of interest in this study, and because forecasting TAI poses unique challenges that require a technical mapping of the space, only a novel technique will be suitable for this problem. To these ends we propose the judgmental distillation mapping (JDM) technique.

The purpose of this technique is not only to develop a scenario map of the plausible technologies that could lead to DTAI or RTAI, like in SNM, but also to generate probabilistic forecasts for each of these scenarios/technologies. No previous techniques have been able to combine scenario planning or scenario mapping with rigorous quantitative forecasts. Inclusion of the rigorous quantitative element is a particularly challenging task because there is so much uncertainty surrounding these technologies, we by default assume them to be five years or more away which is beyond the horizon of reasonable forecasts (Tetlock and Gardner 2016). It is also very challenging because many of these technologies cannot be defined with enough precision to be suitable as forecasting targets (Dafoe 2018).

Yet, forecasts in this manner are necessary for planning and decision making with regard to TAI systems, so some method is necessary. JDM tackles the problem by first generating a map of the different plausible paths forward to TAI/DTAI/RTAI. In order to develop the map, rounds of expert interviews, similar to those described in section 3.2.5.1, are utilized. However, we think of the interview rounds here more as an iterative process like the Delphi rather than as FCMs. Here we proposed three distinct rounds informed by preliminary forecasts, but the process is flexible and can be adapted based on specific circumstances[76]. In general, the first step focuses on identifying the common concepts and definitions that may lead to the development of plausible scenarios. For this step we draw from techniques of the intuitive logics school of scenario planning to develop a cognitive map. The second step involves using the interviews to clarify meanings and to further refine the cognitive map from the scenarios. During this step some participants are presented with a working cognitive map and asked for their opinion, criticism and suggestions. The final step consists of soliciting participants to identify the causal relationships that they believe are present between the concepts in the cognitive map developed from the previous step.

Experts for this process were selected based on the same criteria as for the structured interviews, however, for these interviews experts were targeted more specifically based on unique research foci's relevance to elements of the map that needed refinement or verification. The iterative interviews used for this study were inspired by the iterative rounds of the Delphi process, and we do not suggest that interviews are the only solution for this iterative portion of JDM. Questionnaires are another alternative, and, in many situations they may be more useful. However, for this study it was necessary to identify experts and to establish positive rapport in order to conduct interviews. Because these interviews were more targeted than the previous structured interviews[77], approaching the experts at the conferences, after their speaking engagements, was found to be a more effective way to obtain interviews. Obtaining expert opinion may not be as challenging for organizations wanting to utilize this process to create forecasts that either have

[76] For example, if an organization had already performed the method for generating a previous forecast, then it is possible that only two, or perhaps even just one, round of expert interviews would be necessary to refine the previous map before proceeding to the forecasting step.

[77] The holistic framework, described at length in the following section, requires technical expertise on the part of the forecaster in order to make subjective decisions that can significantly impact the forecasts, such as which experts to include. Usually a team is recommended due to this aspect, however, this was not possible for demonstration purposes.

substantial recognition, such as governments and think tanks, or have experts in-house that they can utilize.

The resulting judgmental distillation map differs greatly from both the scenario network map created in this study and the roadmap of the HLAI technical landscape presented by Adams et al. (Adams et al. 2012). With respect to the former it is less intricate and complex while with respect to the latter it is more complex (being a directed graph as opposed to the grid created by Adams et al.). With respect to each, the JDM map will also differ in that it will include probabilistic quantitative forecasts. While the resulting JDM and SNM maps cannot be compared quantitatively, they will be compared qualitatively to determine their utility and value. Based on such analysis we will recommend uses for each where they may be helpful for others working in AI strategy.

Figure 6 depicts an example JDM process. On the left it can be seen that an input mapping and input scenarios are necessary to build on during the iterative process of questionnaires and interviews. The first step of the process is the generation of the scenarios/technological components of the map. This has been described in the previous portion of this section. On the right side of Figure 6, in the JDM process, can be seen a circle denoting Monte Carlo simulation. Monte Carlo simulation is the mechanism that is used to take the distributional forecasts from each of the various external nodes and aggregate them to create forecasts for the internal layers of the scenario map.

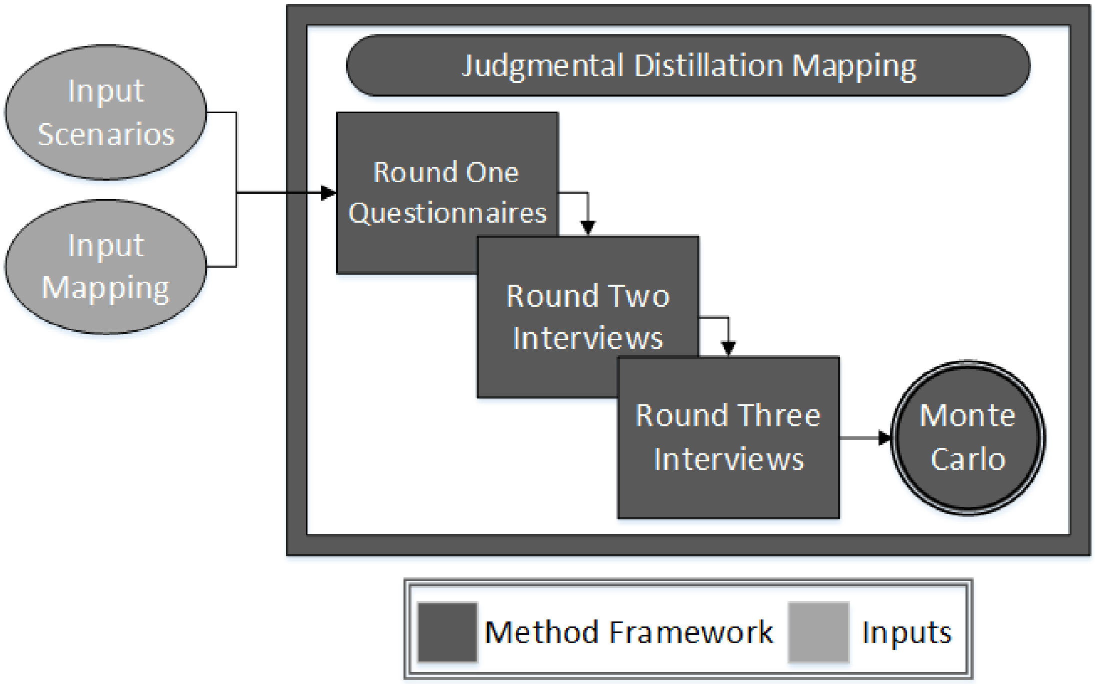


Figure 6: a flow chart depicting a hypothetical judgmental distillation mapping (JDM) process. Questionnaires and interviews can be used interchangeably and involve an iterative process that builds on the input mapping and scenarios. Given probability distributions for the external nodes, probabilistic forecasts for the internal nodes can be generated using Monte Carlo simulation.

An example scenario map is shown in Figure 7. This map was taken from the first round of interviews and is simply meant as an example for understanding the method rather than being an example of the anticipated results from this technique. The nodes in the graph (or map) are representative of technologies (or scenarios). The scenarios are connected by arrows denoting the casual relationship between the nodes. The external nodes are technologies which are near-term plausible and the targets of ongoing research. They may be ambiguously defined for the purpose of forecasting, but decomposition can be used to address this (so they are consistent with the desiderata for forecasting targets; Dafoe 2018). However, these technologies are assumed to be tractably forecastable so that for the purpose of the JDM process they can be the focus of the iterative interviews or questionnaires; decomposition can be completed independently. These tractably forecastable nodes are white in Figure 7.

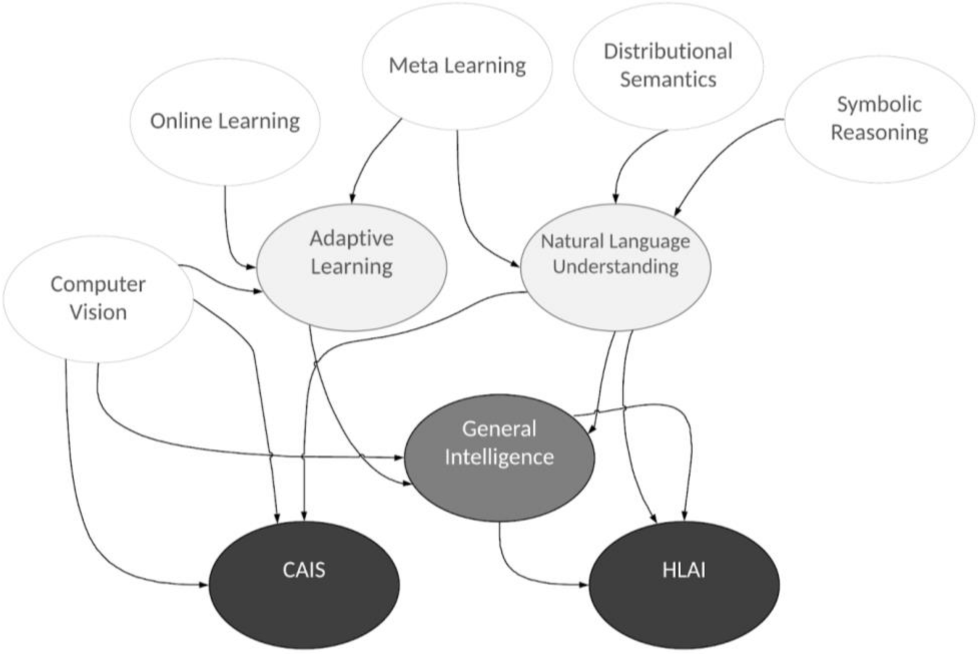


Figure 7: An example of a scenario map from the beginning of the first round of the JDM process. The nodes are representative of the scenario/technologies that comprise the map. The white nodes are external nodes, which are technologies that are on the near-term horizon and are assumed to be tractably forecastable. The light grey nodes represent the first layer of the internal nodes. The darker grey and the darkest nodes represent the end states of the scenario map which correspond to DTAI and RTAI.

The inner nodes of the graph are colored in different shades of grey. The first internal layer is light grey, and these nodes represent technologies that may occur as a direct result of the development of the near-term plausible, tractably forecastable technologies (i.e., the white nodes). These light grey nodes can be thought of as next generation technologies. Probabilistic forecasts for these technologies can be computed by simple Monte Carlo simulation using all of the nodes with inbound causal relationships (given weights for each of the causal relationships). The darker grey nodes represent more distant AI technologies which experts anticipate being possible following the development of the next generation technologies. Probabilistic forecasts for these technologies can be computed in the same manner as the next generation technologies, using a Monte Carlo simulation. The final layer contains the dark nodes which represent the end-state AI technologies (i.e., RTAI) such as CAIS or notions of AGI. Forecasts for this layer would be computed in the same manner. Earlier nodes in the scenario map can correspond to either TAI or DTAI, but it is likely that only the final layer would bring about transformative societal change which would be considered to be RTAI.

The external nodes may be considered to be tractably forecastable as they are within a plausible forecasting horizon, however, they may not be practically forecastable as they likely will not satisfy the desiderata for good forecasting targets (Dafoe 2018). Thus, it may be necessary to decompose these technologies into forecasting targets that are more practical. Figure 8 takes one of the external nodes from the scenario map in Figure 7 and demonstrates how this could be decomposed into more practically forecastable forecasting targets. Using this decomposition, we are able to get probability distributions for these forecasting targets using some method of expert elicitation, and then we would be able to use Monte Carlo simulation to approximate the distribution for the external nodes of the decomposed node. This represents another layer of nodes in the map that is hidden for the purpose of eliciting the map[78]. In order to determine how best to decompose the external nodes would likely require a separate round of expert interviews or questionnaires[79].

[78] It is easier to get constructive feedback on the map from experts when granularity is minimized so that information overload is avoided. This is a particular danger because the map is a visual artifact, and it should ideally be understood quickly and intuitively, without too much cognitive burden. It may even be useful to remove portions of the map other than external node decompositions if future maps become substantially more complex than the example used here for demonstration. We also note that in early rounds, eliminating some detail can be useful because it can have the effect of eliciting differing opinions from different experts (i.e., one expert may see a gap in the map another does not, or experts may see different technologies being able to fill the same gap in the map).

[79] We recommend using the multiple rounds of interviews and questionnaires so that a minimal burden is placed on individual experts' time. When considering this decomposition, we suggest that the level of expertise necessary for informing how best to go about this may not need to be as extensive as the level of expertise necessary for eliciting direct input on the technologies comprising the map or the weights of the causal relationships between them.

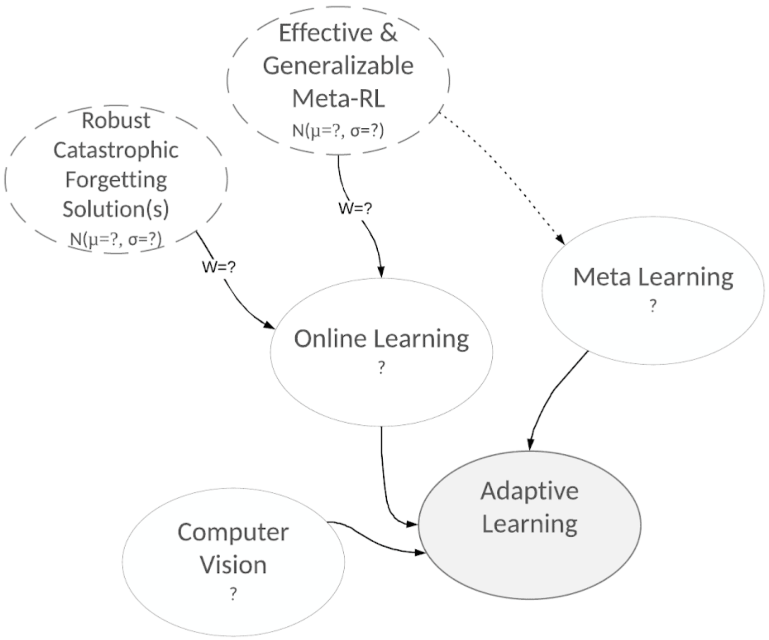


Figure 8: this depicts a decomposition of the "Online Learning" node in the scenario map. It also includes unknown distributions representative of the forecasts that would be needed for these nodes in order to generate a forecast for the focus node.

Monte Carlo simulation has been described as the technique to be utilized in order to generate the probabilistic forecasts for the internal nodes in the scenario map for the JDM process. Unlike many forecasting techniques, here, probabilistic forecasts refer to distributions representative of the timelines over which the technologies may be developed. Moreover, Monte Carlo simulation is a broad class of stochastic techniques for modeling a wide variety of applications, thus the specific details of the implementation here are likely not obvious even to those familiar with these techniques. JDM is presented here for demonstrative purposes, and the specific details of the implementation are also not critical.

We are concerned with modeling probabilities, and the Monte Carlo simulation involves taking probability distributions for the input forecasts and using weights which must be assigned for each of the causal relationships between the nodes to compute an output distribution for the upstream (or output) node. Monte Carlo techniques are a computational technique, which, when used in the context of probability distributions refer to the use of random sampling from a given distribution. Here, the Monte Carlo simulation involves randomly sampling from each of the input distributions and taking the maximum forecast to be the time at which all of the requisite technologies are available enabling the upstream technology. For each group of samples, this completion time of the latest technology would be added to the distribution for the upstream technology. With a large number of samples (e.g., n = 1,000 or n = 10,000) this gives a rough approximation for the upstream technologies. Figure 7 shows the nodes related to online learning and further depicts a hypothetical decomposition of the nodes.

We do not propose this to be the optimal technique for computing approximations for the upstream technologies, however. Other considerations are likely also important. For example, using the causal weights can enable modeling technology development with the possibility of discontinuous progress. There are different approaches to using the weights for this purpose, and one proposed technique is demonstrated here. This utilizes the same principle previously described, but, based on the normalized weights one of the input distribution's samples would be selected, and the value of that sample would be used for the output[80].

5.2 A Holistic Forecasting Framework

JDM is an essential component of the holistic forecasting framework that is proposed here for forecasting TAI. However, while a specific holistic forecasting framework is proposed here, it is important to remember that this is just one example of what would be considered a holistic forecasting framework. We define a holistic forecasting framework to be a technology forecasting framework that utilizes a combination of judgmental forecasting methods, statistical forecasting

[80] The implementation can be found in JDM.py at https://github.com/rossgritz/research/.

methods and scenario analysis techniques in a manner that enables forecasts to be made cyclically, building on knowledge from the earlier forecasts. Such methods are likely best suited for forecasting TAI/DTAI/RTAI due to the complex nature of the forecasting target. To be clear, the two criteria of a holistic forecasting framework are:

1. The framework utilizes a combination of judgmental, statistical or scenario analysis techniques.
2. The framework enables forecasts to be made cyclically building on previous forecasts.

The holistic forecasting framework proposed here involves judgmental forecasting techniques, statistical forecasting techniques and scenario analysis techniques. It also provides for the cyclic generation of forecasts, which enables subsequent forecasts to build on knowledge and insights gained from the expert opinions elicited in previous forecasts. Figure 9 depicts this proposed framework, including the JDM process as the central component. In it, the light grey elements are input forecasts. The darker grey elements are the JDM component, and the darkest grey elements are the strategic planning component. JDM is depicted as a self-contained unit, although this is suboptimal if JDM is implemented within the framework because group expert elicitation techniques can be very beneficial for eliciting the forecasts necessary for Monte Carlo simulation (and such techniques are preferred having been demonstrated superior for such purposes). The rectangular components are input forecasts or contributing forecasts. The circular elements are intended to represent actionable forecasts that can be used by decision makers. The oval elements are aggregated results from input forecasts.

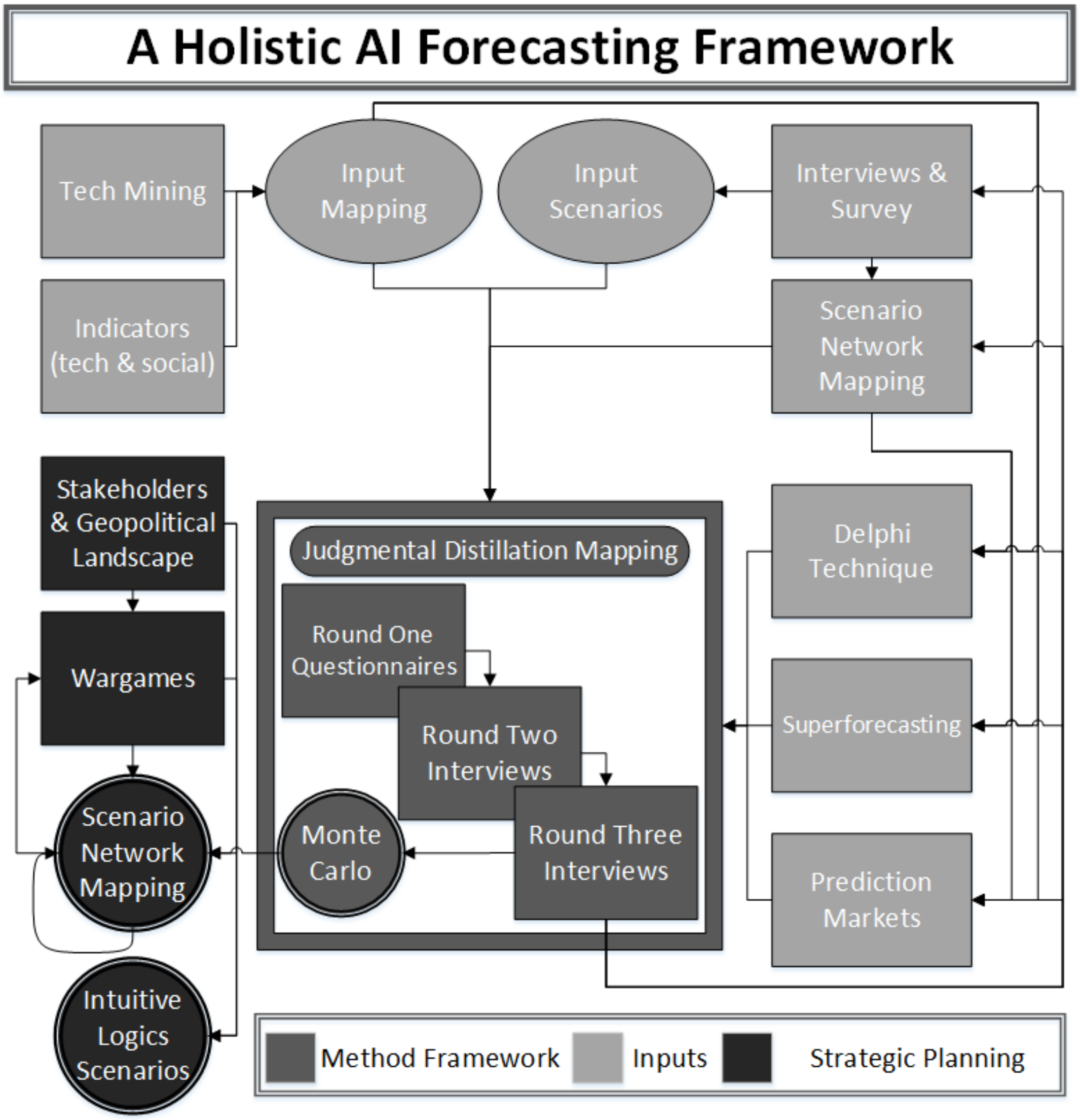


Figure 9: The proposed holistic forecasting framework. This is not necessarily an optimal holistic forecasting framework, however, it does meet the criteria of a holistic forecasting framework proposed here.

The framework encompasses a comprehensive forecasting effort involving numerous techniques which begins with no knowledge relevant to TAI and results in a strategic planning cycle coupled with an ongoing forecasting cycle. The framework proposed here is a rather involved process, and results in the maximal aid to decision makers. Consequently, it may be an extreme alternative for small organizations simply trying to create strategic plans to drive research efforts and revenue. It is best suited for large organizations such as governments, major AI research laboratories or other well-funded research laboratories, and well-funded non-profit organizations concerned with the governance of AI and ensuring the safe and beneficial use of AI. The strategic planning block represents a cycle in itself and could be elaborated on much more extensively in future work. This block is also a critical element, because the actionable forecasts from JDM are likely not well suited for dissemination to decision makers who are not proficient with the underlying AI technologies being forecast.

The input forecasts on the left side of the framework in Figure 9 represent statistical or data driven forecasts (i.e., tech mining and tech and social indicators). These are considered to be aggregated into an input mapping. The initial input forecasts on the right side represent judgmental forecasting techniques (i.e., surveys and SNM). SNM is useful here as an input as SNM alone is not suitable for quantitative forecasts[81] because 1) no method exists for using it for this purpose and 2) it is anticipated to be too granular to use for practical quantitative forecasts. SNM and surveys (also structured interviews) are thought to be aggregated here into scenarios. The input mapping and scenarios can be used to generate an initial scenario map that can then be used for JDM.

During JDM the input mapping is iterated over during rounds of interviews or questionnaires, as described in the previous section. The results of these iterative rounds of interviews have two outputs 1) a map for which forecasts are elicited and which is used to generate quantitative forecasts 2) a scenario map that qualitatively can be used to inform the next rounds of the holistic framework and the JDM process. The first of these requires operationalizing the decomposed external nodes from the map for use with one of a variety of strong quantitative

[81] Qualitatively, SNM has substantial value that JDM and the holistic framework do not offer, such as acting as a research roadmap.

forecasting techniques. For this framework, we propose either the Delphi technique, prediction markets or forecasting tournaments (in the style described by Tetlock and Gardner (2016)). These methods are recommended because the literature review found them to be the strongest methods for quantitative forecasting. The second of these would require using the qualitative results as input for future SNM workshops, surveys and structured interviews. This feature satisfies the 2nd criterion for a holistic forecasting framework: that the framework enables the generation of cyclic forecasts.

The output of the JDM process is a primarily quantitative map of timelines for likely technologies. This becomes well suited for exploring more granularly with a SNM workshop run by a diverse group involving technical experts, AI strategy or policy researchers, and perhaps economists, philosophers or experts from other relevant domains. The resulting scenario network map would be much more valuable for presenting to decision makers because the best possible quantitative forecasts would be combined with likely scenarios and paths for how these technologies might be reached, their impacts on society and the geopolitical dynamics surrounding their development. At this point the results of this SNM process can be used for simulation games, or wargames (Perla 2011), which are effective both for training and for exploring possible futures (Avin et al. 2020). Separately, the results of this SNM workshop can be used for generating intuitive logics scenarios which might be more effective at persuading decision makers to make the decisions that the scenario mapping techniques point to. This scenario planning technique is commonly used for simplifying complex forecasts to make them more palatable for decision makers (Roper et al. 2011)

5.3 Demonstration of JDM and the Holistic Framework

As noted, conducting JDM within the holistic forecasting framework to generate high-fidelity forecasts is a resource intensive effort, and one that is beyond the scope of this study. While the study does center around the holistic forecasting framework and JDM, these were not mentioned explicitly in the original proposal. A demonstration of FCM in the context of forecasting TAI was included, but no notion of a larger framework was. Rather than complete the FCM demonstration, the new method, JDM, was proposed and is demonstrated for this study. The entire holistic framework will not be demonstrated due to the resources required; however, many components will be demonstrated and, due to its significant implications, much discussion regarding it will be included.

To demonstrate JDM three rounds of interviews were used. The first round of interviews was focused on the development of an initial map. These interviews were conducted in late 2018 and early 2019. During these interviews no map was displayed to participants, and most questions focused on experts' respective areas of expertise. The data was analyzed as it was collected and, at the end of this process, a final analysis of the aggregated data was conducted and an initial scenario map was generated. The second round involved, after explicitly discussing the mapping goal of the project, displaying the initial map to the interviewees and eliciting feedback on gaps in the map and the suitability of different components. Again, the interviews were tailored to the participants and focused primarily on their respective areas of expertise. The third round of interviews was actually informal and tested a technique of eliciting expert opinion that simply involved discussing specific topics of interest with experts during question and answer poster sessions at conferences. These were not considered interviews, rather just public question and answer sessions[82]. Such conversations would involve discussion of experts' broader research focus, often an external node in the working judgmental distillation map, and the different subproblems to solve to realize their broader goal. This effectively amounted to eliciting information for forecasting target decomposition. This third round of informal discussions also offered an opportunity to elicit information regarding the weights between different nodes in the judgmental distillation map. This was also done by simply asking about the proportion of contributions for different components of the map. This was frequently challenging because, without showing the working map, many times unincorporated elements would be discussed. Efforts were made to frame anything outside the existing map in terms consistent with the map components and causal relationships for which weights were being elicited. The use of this alternate approach had both strengths and benefits, which are discussed in the results and discussion sections.

For all JDM interviews experts were selected based on the same criteria as used for the earlier structured interviews, but a large degree of subjective judgement on the part of the principal investigator was also utilized. This is a particularly nuanced element of the JDM process; it assumes that those orchestrating the forecasting effort are experts in forecasting with substantial expertise in technical AI and AGI development. This makes operationalizing the technique even more challenging. Furthermore, we suggest that teams of such experts are used for expert selection and survey design, so that there must be a consensus for selecting experts at different stages in the process and consensus for the specific knowledge to elicit from each expert. Doing this with a team of forecasting experts with technical AI expertise makes the operationalization of the

[82] Only handwritten notes were taken during these public question and answer sessions.

technique even more challenging, but we suggest that the results are unprecedented for such complex topics and thus worth the extreme challenges.

The holistic framework is demonstrated here only inasmuch as some of the previously mentioned forecasting technique demonstrations contribute to the framework. Because of the chronology of the different components of the study conducted to demonstrate the various techniques, the input forecasts for the JDM process were not all used. Results from the structured interviews were used for informing the first round of the JDM process, but the initial JDM map was developed based on the first round of JDM interviews. Results from the first SNM became available after the first round of the JDM interviews and were used as an input for the second round of the JDM process having a nontrivial impact on the resulting map. The extrapolation demonstration was used as an input for the second round of the JDM process (it was not available before this), and it elicited interesting results. However, these results were not very significant with regard to the JDM process, but still have implications for future forecasts using some variation of the proposed holistic framework. Other methods, such as the survey and the tech mining demonstration were not used significantly in the holistic framework and were only primarily used for demonstrating the independent techniques. Yet, the demonstrations of these techniques were conducted prior to the beginning of the JDM process, so the results were sometimes mentioned as relevant during interviews, and some information gathered through conducting these processes was tacitly used to develop the interview questions for the early rounds of JDM.

5.4 Possible Alternate Frameworks

As explicitly noted in the beginning of this chapter, the framework proposed here is an example of a holistic forecasting framework. Given the criteria proposed for holistic forecasting frameworks, there are many different possible ways in which the methods discussed in this study (and potentially other methods as well) could be combined to satisfy them. However, only one public technique has been identified that meets these criteria. This is discussed below.

Tetlock has proposed a technique he calls full-inference-cycle tournaments[83] (Tetlock 2017). This technique meets the two criteria for a holistic forecasting framework. Briefly, it involves four different stages of a forecasting tournament that is intended to operate cyclically. The first stage involves scenario generation (called "gists" in the proposal, perhaps because Tetlock has previously discussed a distaste for scenario analysis (Tetlock 2006)). The second round involves converting these gists into forecastable targets, possibly using decomposition. Notably, each of these rounds involves a team of forecasters with no domain area expertise, similar to Tetlock's previous work utilizing forecasting tournaments (Tetlock and Gardner 2016). The third round involves actually forecasting the targets generated in the previous stage. The final stage involves conducting a "post-mortem" wherein the results of the forecasts are analyzed and relevant knowledge is summarized for passing on to the next full-inference-cycle, which begins again with the first stage of generating gists.

FCITs are presented as an alternative formulation[84] of the notion of a holistic forecasting framework presented here. It is significant in that it demonstrates that the notion of a holistic framework proposed here is consistent with some of the latest work from leaders in the field. The

[83] The concept discussed here was proposed to the Intelligence Advanced Research Planning Activity in 2017 but was unfunded. The original proposal also focused on the use of the technique for forecasting geopolitical forecasting targets, similar to Tetlock's other work. With modifications it may be a suitable forecasting framework for forecasting AI progress.

[84] It is possible that another holistic forecasting framework is also beginning to be used by a think tank in Washington DC (as this is not public knowledge, we do not identify the think tank). This think tank has been developing a large-scale dataset for use in mapping and forecasting AI progress for a year or greater. However, more recently (and roughly a year after the publication of the paper on the holistic forecasting framework presented here; Gruetzemacher 2019a), they have begun a forecasting tournament effort for technology forecasting. It is likely that the information extracted from the tech mining efforts will be used in some way to inform the questions or foci of the questions explored by Foretell. If true, and if the results from the forecasting tournament discussions or forecasts were used to inform future rounds of forecasting – i.e. in a cyclic manner – then this would also constitute another example of a holistic forecasting framework (and one in the context of AI forecasting).

framework proposed here is intended to be a contribution to the existing literature, but it is also hoped that the notion of a holistic forecasting framework, and the existence of multiple examples of such frameworks, can also enable forecasting researchers to consider the virtues of more holistic thinking for future research. It is important to consider that both of the examples here require substantial resources, likely requiring a minimum of several hundred thousand dollars for a single iteration of each framework[85]. Despite this, for topics of grave importance and extreme complexity, such as TAI/DTAI/RTAI, large investments such as this could pay very large dividends if successful.

[85] The same would likely be true for the think tank effort if it is in fact an example of a holistic forecasting framework.

6 RESULTS

This chapter covers the results of the various demonstrations conducted as part of this dissertation. This includes the results for six of the seven techniques discussed in Chapter 4, the only exclusion being the results from the Delphi which are discussed separately in Chapter 7.

6.1 Survey

This section explores the results of the survey data and analysis. The survey included 13 major questions. Here we break these questions down into four primary categories: calibration questions (Q1-Q4[86]), fixed years forecasts (Q5[87]), fixed probabilities forecasts (Q6-Q10[88]) and compute related questions (Q11-Q13[89]). The results from each of these sections are discussed in separate subsections. The full results can be found in Appendix C[90].

6.1.1 Calibration Questions

The first questions in the survey were the calibration questions. As described in Chapter 4, these questions were created with forecasts that attempted to forecast targets over the following four years. Consequently, not all of these questions are able to be evaluated at the time of this writing. Because of this, a novel technique is proposed to make use of this data that is referred to here as naïve calibration. The naïve calibration technique relies on the assumption that extra-confident

[86] These are the question numbers that were used for analysis purposes; i.e. in the Jupyter notebooks and R code used for the analysis, which can be found at https://github.com/rossgritz/labor-displacement-survey-2018, these are the question numbers that were used. These questions correspond to questions 3-6 in Appendix A and Appendix B.

[87] Q5 corresponds with the questions 7, 8 and 9 in Appendix A and Appendix B.

[88] These questions correspond to questions 11, 13, 14, 16 and 17 in Appendix A and Appendix B (for Q6-Q10, respectively).

[89] Q11-Q13 correspond to questions 19, 20 and 21 in Appendix A and Appendix B.

[90] The question numbering in Appendix C is consistent with this numbering. Appendix A and Appendix B use a different numbering that was necessary for the survey instrument but did not represent independent questions of interest in this study.

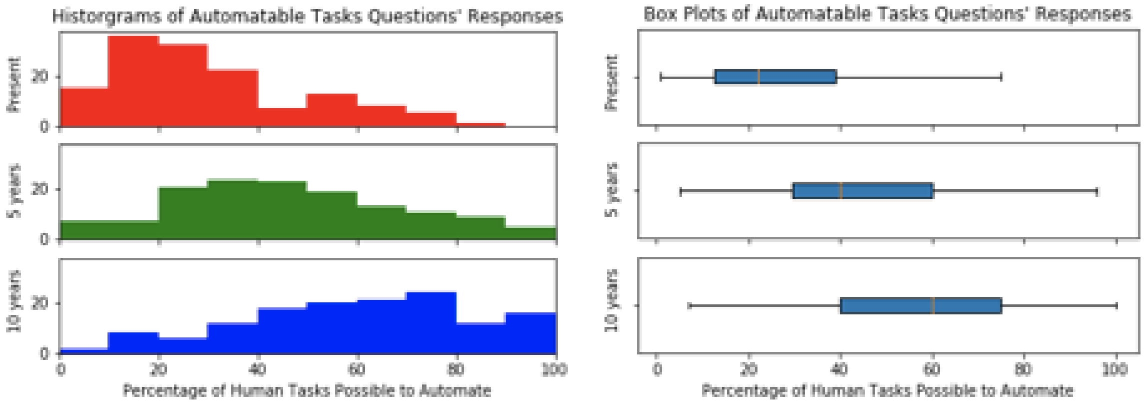
forecasts, defined as forecasts with complete certainty (i.e., forecasts of 0% or 100% probability), are fundamentally flawed and thus are an indicator of poor forecasting calibration. This characterization follows from the fact that none of the participants could know the resolution of any of these questions with absolute certainty, and thus, forecasts from persons forecasting any one of the events with absolute certainty were poorly calibrated. Even if all of the calibration questions could be evaluated, the sample is too small to be used reliably to calibrate forecasts. However, using the naive calibration technique proposed, such fundamental flaws are not acceptable at any frequency, whether from a large or small sample, so sample size is irrelevant. This is a novel method, and it relies on a smaller number of calibration questions making it better suited for surveys due to challenges posed by survey fatigue[91] (Creswell 2014). The results of the naive calibration are reported in more detail in a subsequent subsubsection, following the presentation of the results on fixed probabilities forecasts in which distributions were elicited from experts and this new technique was applied and evaluated (specifically, section 6.1.3.3).

6.1.2 Fixed Years Forecasts

The fixed years forecasts involved a single question that asked for the percentage of human tasks that would be feasibly automatable at present and at fixed times in the future (5 years and 10 years). The results from this question were straightforward and did not require any specialized processing for visualization; descriptive statistical analysis was sufficient. Figure 10 depicts these results. Histograms can be seen in Figure 10a and box plots can be seen in Figure 10b. The median forecast for the portion of human tasks that it is feasible to automate currently, using AI technology, is 22%. For five and ten years, this number rises to 40% and 60%, respectively. The majority of mass in the distributions depicted in Figure 9a can be seen to shift from the current estimates of

[91] Survey fatigue refers to participants' tendency to become uninterested, and consequently, to perform poorly, on long surveys.

automatable tasks to those estimates in 10 years. This may suggest a consensus among practitioners regarding the increase of the automatability of tasks using AI technology over the next decade.



(a) (b)

Figure 10: a) histograms for the portion of human tasks currently automatable using AI and b) box plots for the portion of human tasks currently automatable using AI.

6.1.3 Fixed Probabilities Forecasts

The survey results for Q6-Q10 were aggregated for visualization and analysis of these five levels of extreme labor displacement from AI (i.e. the fixed probabilities forecasts). These five forecasts are presented in different visualizations in the following subsubsection. Descriptive statistics for each of the five forecasts are presented in Table 5. For the purposes of analysis and discussion we are primarily concerned with the median 50% probability forecasts. The median forecast for narrow systems capable of accomplishing 50% of human tasks is 10 years, which varies slightly from the forecast of 60% of tasks being automatable in 10 years in the fixed years questions. This inconsistency may be attributable to some cognitive bias such as round number bias or framing effects of the order in which the questions were asked (Fraser-Mackenzie et al. 2015; Kahneman 2011). Moreover, round number bias seems likely to be a challenge for many of the forecasts as demonstrated by the fact that four out of five of the 50% median forecasts are a factor of five. This is certainly something that should be considered by researchers in eliciting opinion for future work.

Table 5: Simple Descriptive Statistics for Q6-Q10 Results (in years)

		System Type				
	Probability	50% Narrow	90% Narrow	90% Broad	99% Narrow	99% Broad
Median	10%	5	10	15	25	30
	50%	10	25	32.5	50	50
	90%	28.5	50	60	99	100
Mean	10%	8.567	33.65	27.88	41.79	57.75
	50%	21.99	65.24	60.58	78.10	106.2
	90%	75.44	184.2	146.4	174.8	218.7
Median Aggregate Forecast	10%	3.357	10.15	16.07	22.09	28.31
	50%	11.63	24.76	33.16	50.42	56.46
	90%	28.36	49.42	59.54	96.63	99.13

Perhaps of most interest among the results presented in Table 5 are the 50% median forecasts for the four extreme labor-displacing AI scenarios. These indicate that AI practitioners anticipate narrow and broadly capable systems to be capable of completing 90% of human tasks in 25 years and 32.5 years, respectively. They further indicate that for both narrow and broadly capable systems practitioners anticipate this being likely in 50 years. Thus, for this most extreme case there was no difference between the narrow system and broadly capable system forecasts.

To visualize the results from these questions we use the method of least squares to fit each respondent's data to a gamma distribution, then we aggregate these distributions using the median aggregation technique. This is discussed in the following subsubsection. Multivariate regression involving two dependent variables is then used to test for a difference between the HLAI researchers and the ICML/IJCAI researchers.

These were the core questions of the survey. They asked for participants' forecasts in years of when fixed portions of human tasks would be automatable from AI. These questions focused on extreme labor displacement scenarios that would constitute DTAI and RTAI. These questions were also used for analyzing the primary question explored in the study: is there a difference between forecasts from HLAI researchers and mainstream AI researchers.

6.1.3.1 Visualization and Comparisons

In order to visualize the results of these forecasts the three data points for each individual forecast were fitted to a gamma distribution following the work of Grace et al. (2018). Specifically, the three quantiles were fit to the cumulative distribution function (CDF) for the gamma distribution using least squares. The gamma distribution can be parameterized with two parameters, denoted α and β. Equation 1 depicts the gamma distribution as a function of α and β.

$$F(x;\alpha,\beta) = \int_0^x f(u;\alpha,\beta)du = \frac{\gamma(\alpha,\beta x)}{\Gamma(\alpha)} \qquad (1)$$

Above, $\gamma(\alpha, \beta x)$ and Γ are the lower incomplete gamma function and the gamma function, respectively. The gamma function is useful for different applications in probability, statistics and combinatorics. Rearranging Equation 1 yields Equation 2:

$$F(\boldsymbol{y};\ \alpha, \beta) \ = \ \beta\gamma^{-1}(\alpha, \boldsymbol{y}) \tag{2}$$

The aggregation technique used was a simple median-based technique (Hora et al. 2013). This simply involves aggregating each quantile and taking the median values. The vector of quantile values is denoted as x, and the vector y is set equal to the fixed quantile probabilities that correspond to each forecast, i.e., [0.1, 0.5, 0.9]. Using these vectors, and applying the method of least squares to minimize Equation 3 we are able to approximate α and β.

$$\boldsymbol{x} \ - \ \beta\gamma^{-1}(\alpha, \boldsymbol{y}) \tag{3}$$

Equations 1 through 3 were used to create each of the remaining figures in this section depicting the results as aggregated CDFs. Figure 11 depicts the results for each of the five fixed probabilities questions in a single plot. In it, slight differences can be seen between the narrow and broadly capable system forecasts for each of the different questions. This difference is more significant in the systems capable of accomplishing 90% of human tasks as opposed to the more extreme labor displacement scenario systems. However, because these distributions are aggregated parameterized probability distributions there is not an appropriate test for testing the statistical significance of these differences.

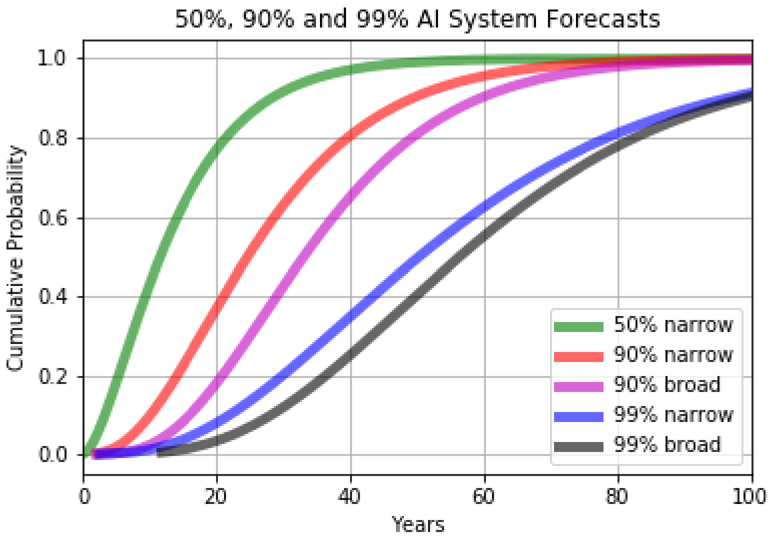


Figure 11: The results for each of the five fixed probabilities forecasts (Q6-Q10).

Figure 12 depicts a comparison of the forecasts from the different conference locations for a 99% broadly capable system. Here, bootstrapping (n = 10,000; Effron and Tibshirani 1994) has been used to show 95% confidence intervals for each of the aggregated distributions. This highlights the significance of the difference between forecasts for each of these test groups. Using the multivariate model, detailed in the following pages, this difference is found to be highly statistically significant[92], as would be expected based on the visualization of the results depicted in the Figure.

[92] This was possible because we were evaluating the statistical difference between two groups and we did so using a multivariate model. The aggregate distributions themselves were not evaluated statistically.

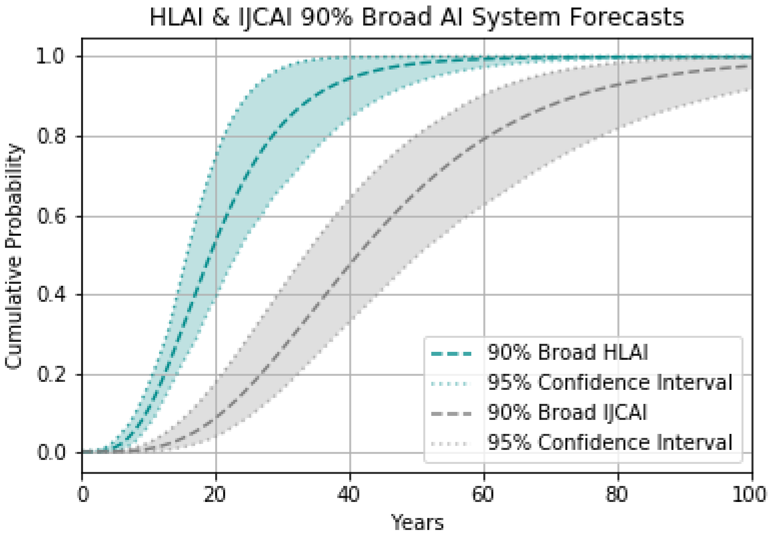


Figure 12: A comparison of the 99% broadly capable system forecasts for the two conference locations. 95% confidence intervals are depicted as shaded regions for each aggregate distribution.

Figure 13 also depicts aggregated distributions from the results using bootstrapping (n = 10,000) for showing 95% confidence intervals. In this figure, the comparison is between forecasts for the 90% broadly capable system and the 99% broadly capable system. The 95% confidence intervals used here also clearly depict the significance of the difference between the two forecasts. While this difference was not modeled statistically, however, there is clearly a significant difference.

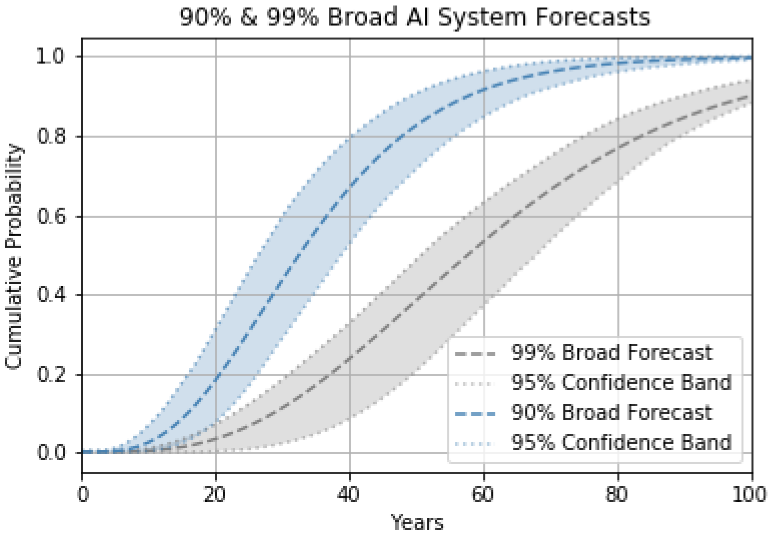


Figure 13: A comparison of the 90% and 99% forecasts for broadly capable systems, including 95% confidence intervals.

Table 6 contains the descriptive statistics for comparing the results from the two different conference locations. The particularly interesting results depicted here involve those for the 10% median forecasts for the HLAI conference attendees. Practitioners' from this group anticipate a 1-in-10 chance of 90% and 99% of tasks being automatable from AI in 10 and 20 years, respectively, for both narrow and broadly capable systems. Further comparison reveals that while there is no difference between narrow and broad forecasts for HLAI practitioners for 90% and 99% systems at a 10% probability, these respondents do expect an increase of five years for developing a broadly capable system beyond narrow systems for each 90% and 99% capable AI systems at 50% probability. However, neither the anticipated difference between narrow systems and broadly capable systems at 50% probability, or the lack of a difference for the 10% probability forecasts, holds for the 90% probability forecasts. This lack of consistency for the tails of the forecast distributions for definitively different types of systems is interesting, and perhaps a consequence of round number bias, but could also be an indicator of increased uncertainty.

Table 6: Comparison of Median Forecasts from Different Conference Locations

		50%	90%		99%	
		narrow	narrow	broad	narrow	broad
10%	HLAI	5	10	10	20	20
	IJCAI	5	15	20	30	46
50%	HLAI	10	15	20	30	35
	IJCAI	12	30	41	50	70
90%	HLAI	20	31	33.5	50	50
	IJCAI	30	60	70	100	102

6.1.3.2 Multivariate Statistical Analysis

Statistical analysis was conducted to assess the effect of the practitioners' conference of attendance as well as other demographic factors. A seemingly unrelated regression (SUR) model was used (Zellner 1962). This is an appropriate model because, if the error terms are uncorrelated, the model simply becomes a set of ordinary least squares (OLS) models (Green 2000). For the model we use two dependent variables, Y_m and Y_u. Y_m is the 50% median forecast dependent variable and Y_u is the uncertainty dependent variable. No known previous work which has elicited data as quantiles for 10%, 50% and 90% has modeled the uncertainty as a dependent variable like this. However, the uncertainty is of significant importance for forecasts like those of concern here that deal with what can be considered deep uncertainty (Marchau et al. 2019). Here, Y_u is set equal to the difference between the 90% median forecast and the 10% median forecast.

Equations 4 through 7 represent the model that was used. Here, F is the raw input value for each forecast as given by respondents. Equation 4 defines the dependent variable Y_m and Equation

5 defines the dependent variable Y$_u$. Because the data is skewed with a fat tail, a log transformation has been used on each of the dependent variables. The full SUR model is shown in Equations 6 and Equation 7, for each of the dependent variables.

$$Y_m = \log\,(F_{50\%}) \tag{4}$$

$$Y_u = \log(F_{90\%} - F_{10\%}) \tag{5}$$

$$Y_m = \beta_0 + \beta_{1:5}X_{role} + \beta_{6:8}X_{region} + \beta_9X_{conference} + \beta_{10}X_{gender} + \varepsilon_{median} \tag{6}$$

$$Y_u = \beta_0 + \beta'_{1:5}X_{role} + \beta'_{6:8}X_{region} + \beta'_9X_{conference} + \beta'_{10}X_{gender} + \varepsilon_{uncertainty} \tag{7}$$

The SUR model was fit for each of the five different extreme labor-displacing AI scenarios using the R package systemfit for approximating simultaneous equations (Henningsen and Hamann 2007). The results are shown in Table 7, with those for Y$_m$ shown in the upper half of this table and those for Y$_u$ shown in the bottom half of the table. The statistically significant p-values in these results are shaded grey. The intensity of the shade of grey corresponds to the different levels of statistical significance[93] depicted in the table (for a range of p <= 0.10 to p <= 0.005). p-values that were found to be statistically significant for both dependent variables are italicized.

[93] The bottom row of the table acts as a legend for this.

Table 7: Results from the SUR Statistical Model

		System Type									
		50% narrow		90% narrow		90% broad		99% narrow		99% broad	
		β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Y_m	Student	0.697	0.093	0.266	0.542	0.503	0.190	0.130	0.755	0.082	0.851
	Academic	0.480	0.248	0.023	0.958	0.312	0.417	-0.050	0.904	-0.087	0.841
	Research	0.729	*0.070*	0.232	0.580	0.382	0.301	-0.085	0.831	-0.095	0.820
	Engineer	0.606	0.153	0.435	0.329	0.517	0.187	-0.152	0.724	-0.107	0.812
	Executive	0.160	0.732	0.518	0.296	0.538	0.216	-0.035	0.942	0.228	0.642
	Europe	-0.041	0.887	0.070	0.817	0.157	0.566	-0.594	0.042	-0.498	0.102
	NA	-0.177	0.578	-0.270	0.426	-0.096	0.753	-0.674	0.041	-0.668	0.052
	Asia	0.240	0.488	0.342	0.356	0.523	0.108	-0.259	0.456	-0.154	0.670
	HLAI	-0.378	*0.061*	-0.676	*0.002*	-0.687	*0.000*	-0.624	*0.002*	-0.774	*0.000*
	Female	-0.392	0.166	-0.061	0.843	-0.490	*0.079*	-0.259	0.379	-0.367	0.233
Y_u	Student	0.877	0.101	0.129	0.808	0.320	0.518	0.227	0.677	0.317	0.570
	Academic	0.610	0.254	-0.014	0.979	0.071	0.887	-0.043	0.937	0.044	0.937
	Research	1.003	*0.053*	0.320	0.532	0.318	0.506	0.381	0.467	0.438	0.414
	Engineer	0.969	0.077	0.309	0.570	0.411	0.416	-0.227	0.687	-0.044	0.939
	Executive	0.602	0.318	0.173	0.775	0.297	0.596	0.152	0.809	0.437	0.488
	Europe	0.067	0.857	0.090	0.808	0.196	0.579	-0.414	0.277	-0.485	0.214
	NA	-0.063	0.877	0.053	0.899	0.062	0.876	-0.375	0.380	-0.528	0.229
	Asia	0.755	0.092	0.927	0.042	0.966	0.023	0.128	0.778	0.051	0.913
	HLAI	-0.527	*0.043*	-0.855	*0.001*	-0.705	*0.004*	-0.624	*0.019*	-0.758	*0.005*
	Female	-0.324	0.372	-0.323	0.389	-0.618	*0.087*	-0.268	0.485	-0.559	0.157
	Significance	0		0.1		0.05		0.01		0.005	

The primary finding here is that the conference location was statistically significant for all of the extreme labor displacing scenarios for each of the dependent variables. For the four most extreme labor displacement scenarios from AI, there was strong statistical significance for both dependent variables. Because the standard error values are negative for all of these cases, this indicates that the forecasts from the practitioners at the HLAI conference were shorter and had lower uncertainty (i.e., higher precision). Statistically significant correlations were observed for some of the other demographic independent variables, such as region of residence and occupational role, but there were no consistent trends among any of these variables for all of the different extreme labor displacement scenarios or for both dependent variables at a level of p <= 0.05.

6.1.3.3 Naive Calibration Results

The naive calibration model utilized the same model shown in Equations 4 through 7, with two small modifications to Equation 6 and Equation 7. These modifications merely involved the addition of one more dummy variable corresponding to whether or not any of the calibration questions had been answered in a fundamentally flawed manner (i.e. with extreme overconfidence). The results of this are depicted in Table 8, which shows that the results changed very little, and the overall strength of the statistical correlation for the four most extreme labor-displacing AI scenarios remained within the same range.

Table 8: Results of the Naive Calibration

		System Type									
		50% narrow		90% narrow		90% broad		99% narrow		99% broad	
		β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
	Student	0.746	0.072	0.292	0.507	0.539	0.164	0.176	0.673	0.134	0.759
	Academic	0.513	0.214	0.040	0.928	0.320	0.408	-0.051	0.903	-0.098	0.823
	Research	0.817	0.042	0.277	0.513	0.429	0.251	-0.031	0.939	-0.043	0.918
	Engineer	0.619	0.141	0.439	0.326	0.527	0.180	-0.142	0.741	-0.097	0.829
	Executive	0.184	0.695	0.532	0.287	0.567	0.196	0.022	0.963	0.292	0.555
Y_m	Europe	0.007	0.982	0.094	0.760	0.195	0.483	-0.541	0.067	-0.436	0.156
	NA	-0.156	0.627	-0.258	0.451	-0.072	0.816	-0.639	0.053	-0.623	0.071
	Asia	0.254	0.461	0.348	0.349	0.535	0.102	-0.247	0.477	-0.141	0.697
	HLAI	-0.317	0.116	-0.645	0.003	-0.654	0.001	-0.586	0.005	-0.737	0.001
	Female	-0.397	0.164	-0.066	0.834	-0.516	0.070	-0.305	0.308	-0.430	0.169
	Calibration	-0.026	0.947	-0.001	0.995	0.067	0.680	0.143	0.434	0.212	0.265
	Student	0.967	0.072	0.204	0.702	0.423	0.390	0.347	0.517	0.455	0.403
	Academic	0.632	0.237	-0.009	0.987	0.054	0.912	-0.074	0.890	0.006	0.991
	Research	1.110	0.033	0.401	0.437	0.416	0.381	0.500	0.332	0.567	0.280
	Engineer	0.991	0.070	0.333	0.540	0.452	0.366	-0.206	0.708	-0.019	0.972
	Executive	0.695	0.253	0.242	0.690	0.408	0.463	0.332	0.591	0.614	0.319
Y_u	Europe	0.154	0.682	0.166	0.658	0.318	0.369	-0.265	0.481	-0.318	0.404
	NA	0.017	0.967	0.106	0.799	0.152	0.699	-0.269	0.521	-0.406	0.343
	Asia	0.787	0.079	0.935	0.040	1.008	0.016	0.156	0.725	0.085	0.850
	HLAI	-0.470	0.073	-0.797	0.003	-0.636	0.009	-0.542	0.039	-0.665	0.013
	Female	-0.389	0.292	-0.387	0.310	-0.739	0.042	-0.425	0.267	-0.739	0.058
	Calibration	0.413	0.421	0.186	0.415	0.371	0.074	0.508	0.031	0.609	0.011
	Significance	0		0.1		0.05		0.010		0.0050	

The effect of the inclusion of the naive calibration dummy variable was only statistically significant for the uncertainty dependent variable for the two 99% labor-displacing AI scenarios. It is also interesting to compare the R_2 values for the models with and without the naive calibration dummy variable. This comparison can be seen from the data depicted in Table 9. For the 99% broadly capable system forecast the naive calibration dummy variable was able to explain roughly 25% more of the variability in the median dependent variable and over twice as much of the variability in the uncertainty dependent variable. Results for the 99% narrow and 90% broadly capable system forecasts were mixed while the results for the 90% narrow and 50% narrow forecasts indicated that the use of the naive calibration dummy variable marginally decreased the explained variability for each of the dependent variables.

Table 9: A Comparison of R_2 Values for the Different Statistical Models

		R_2	
		Original SUR	Naïve Calibration SUR
99% Broadly Capable System	Y_m	0.0751	0.0936
	Y_u	0.0561	0.1143
99% Narrow Systems	Y_m	0.0751	0.0621
	Y_u	0.0561	0.0812
90% Broadly Capable System	Y_m	0.1632	0.1519
	Y_u	0.1307	0.1413
90% Narrow Systems	Y_m	0.0782	0.0651
	Y_u	0.1073	0.0996
50% Narrow Systems	Y_m	0.0612	0.0574
	Y_u	0.0961	0.0839

6.1.4 Compute Questions

The final three questions concerned the implications of computational resources on future progress in AI research. The first two of these (Q11 and Q12) related to the probability of the trend identified by Amodei and Hernandez (2018) continuing[94]. In order to visualize these results, survival analysis was utilized. Survival analysis was considered appropriate for this application because the data concerned the time to failure of an event (i.e., the trend continuing). The survival function is actually the complement of the gamma CDF, and, assuming that the probability of the function continuing at the time of the survey was 100%, the data was fit to the complement of the gamma CDF. The gamma CDF is shown in Equation 8.

$$S(\boldsymbol{y}) = F(1 - \boldsymbol{y}) \tag{8}$$

Setting x = [0.0, 5.0, 10.0] to represent the fixed years and the assumption for the current probability, the dependent variable is represented as the vector y = [y_{present}, $y_{5\text{yr}}$, $y_{10\text{yr}}$]. The assumption mentioned earlier explicitly implies that y_{present} = 100%. Then, the method of least squares is simply used to minimize Equation 9 to approximate α and β (all other details were consistent with how the gamma CDF was fitted for the fixed probabilities questions).

$$\boldsymbol{x} - \beta\gamma^{-1}(\alpha, \boldsymbol{y}) \tag{9}$$

The results of the fitted survival function are shown in Figure 14. The figure depicts a high likelihood of the trend continuing for the next decade, with this likelihood slowly diminishing over the following three decades. The implications for beyond a decade may be a poor approximation as the question only concerned the next 5 and 10 years, but, the amount of confidence and appearance of consensus is a strong indicator of the value that computational resources are perceived to have in current and future AI research.

[94] See Figures A1 and B1.

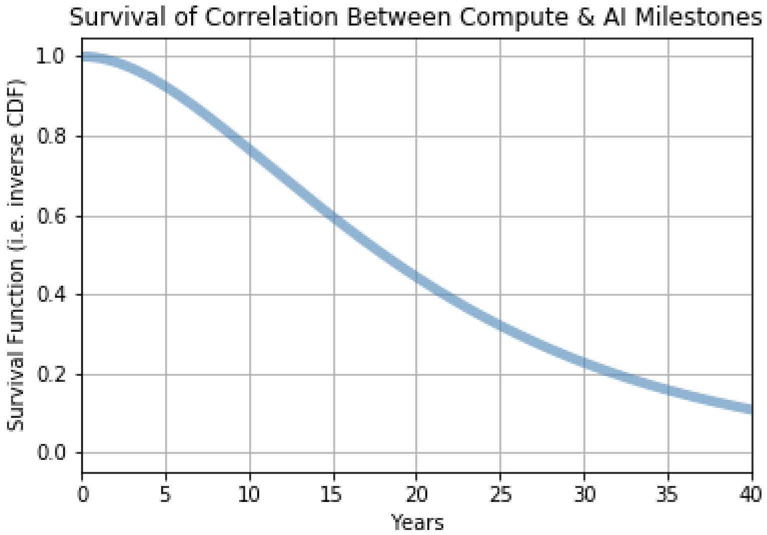


Figure 14: The results from Q11 and Q12 on the probability of the Amodei and Hernandez computational resources forecast continuing. The 5-year and 10-year forecasts were used to fit a survival function (the complement of the gamma CDF) to the data.

The remaining question, Q13, simply asked how much practitioners anticipated progress would increase if given unlimited computational resources. This information was elicited to identify the degree to which computational resources are a bottleneck for cutting edge AI research. The results are depicted in a very straightforward manner as a simple histogram in Figure 15. It shows that the majority of practitioners see progress speeding up if computational resources were no longer a constraint, with over half expecting an increase in research progress by 100% or more. This seems to indicate consensus about computational resources being a bottleneck.

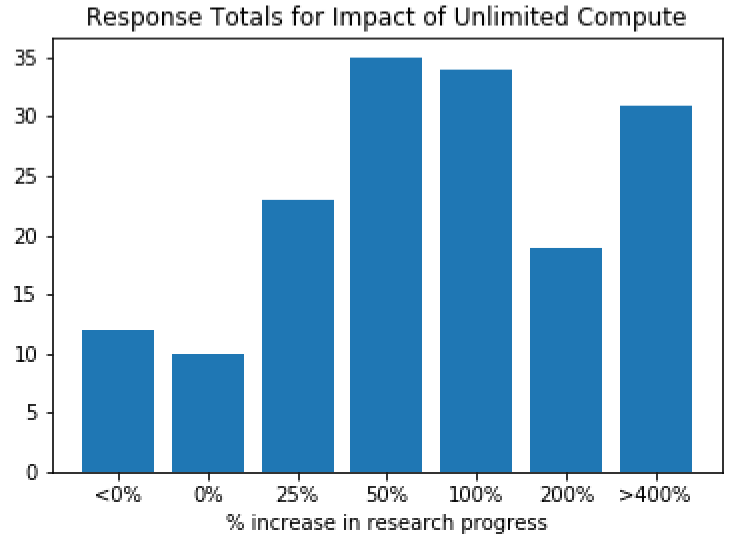


Figure 15: The anticipated increase in progress in AI research if researchers were given unlimited computational resources.

6.2 Tech Mining

The tech mining demonstration was intended as a simple exploratory analysis. For exploratory analysis involving academic abstracts and conference proceedings, two text mining techniques were deemed to be appropriate: clustering and topic modeling. Specifically, k-means clustering and latent Dirichlet allocation (LDA; Blei et al. 2003) were used to identify common themes in the data. T-distributed stochastic neighbor embedding (t-SNE; Maaten and Hinton 2008) was also used for visual analysis of the results of the clustering and topic modeling. However, initially all results were analyzed with descriptive statistics. The k-means and LDA results are of primary interest here and are reported in this section.

We first consider the k-means clustering technique applied to the AI abstract dataset. The clusters have been labeled based on the top words and top titles that resulted from the clustering. This process was used to label all clusters and topics in this portion of the study. For cases where there was some ambiguity in the abstract or conference proceedings' titles, the abstract or full text documents were accessed to determine the best cluster or topic label. Table 10 depicts the top five words and titles for the AI abstract dataset when using k-means. Here, fuzzy logic seems to be a common topic, but probabilistic logic also seems to be a common topic. Thus, because logic is the common topic the label was determined to be “logic.” To be certain, more time could be spent labeling as well as on the analysis, particularly if the k value was not fixed at 10, as it was for our exploratory analysis. This is a tedious and time-consuming technique, but the results often become increasingly valuable with more detailed analysis.

Table 10: Top Words and Titles for k-means on AI Abstract Dataset

Top Words	Top Abstract Titles
fuzzy	Is there a need for fuzzy logic?
logic	Fuzzy-Sets in Approximate Reasoning
control	Possibility theory, probability theory and multiple-valued logics: A clarification
set	Are artificial neural networks black boxes?
rule	An approach to modelling and simulation of uncertain dynamical systems

Figure 16 depicts the t-SNE visualization of the k-means clustering for the AI abstracts that were scraped from Web of Science. The legend shows the colors associated with each of the cluster labels. t-SNE is a method which is commonly used for visualization, and it is well regarded as one of the best methods for visualizing high dimensional spaces by projecting them down to either two or three dimensions (Maaten and Hinton 2008). The suitability of t-SNE for this purpose can be seen in Figure 16. Robots, logic and image processing are topics that are extremes, and far from other topics. Machine learning and artificial intelligence are mixed in the middle with the other topics. T-SNE would be more successful at identifying clear clusters if projecting a dataset of handwritten digits, where there are different classes of data, down to two dimensions because clear groups would be visible. Because there is so much overlap between the clusters the technique is not as effective for clustering, however, it still demonstrates that some of the clusters are relatively strong while other clusters could stand improvement. Logic, image processing and robots may be the most clearly defined clusters as indicated by t-SNE because they are the most isolated. Intelligent agents is another relatively well-defined cluster.

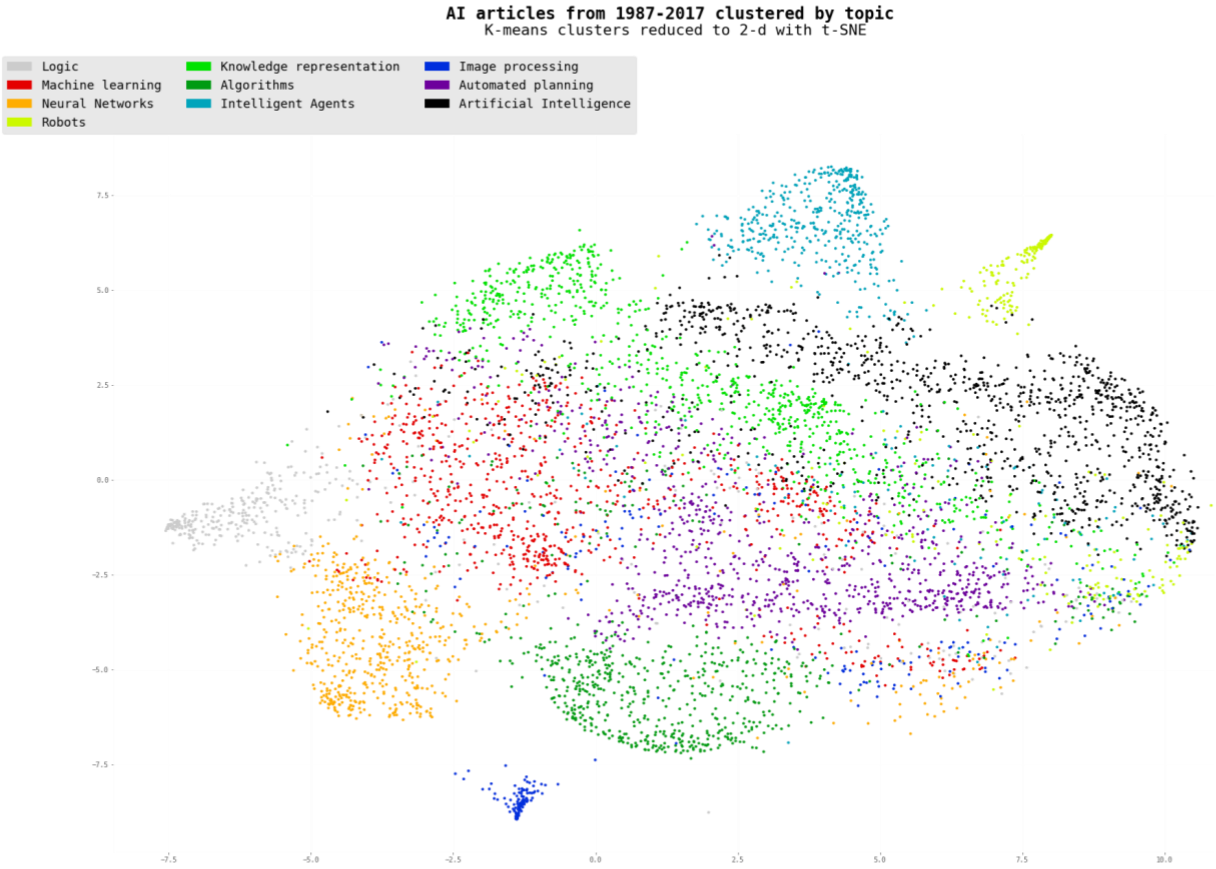


Figure 16: t-SNE used to project the high dimensional dataset down to two dimensions for visualization of the clusters generated by k-means clustering on the AI abstracts from Web of Science.

Because not all of the clusters are well-defined, as indicated by the t-SNE visualization and specific analysis of the content of each cluster, these results suggest that further work could be done to identify clusters that contained stronger signal[95]. k-means was not the only technique assessed for clustering/topic mining; LDA was also used for topic mining to generate topics similar to how k-means was used to identify the most salient topics through clustering. Each technique was used for each of the three datasets. The results from the full text NIPS proceedings are discussed below to illustrate some of the insights from this exploratory analysis. The remaining results are depicted in Appendix D. In general, the k-means results appeared to be superior to the LDA results for identifying salient topics.

This second example of the tech mining results is included because the quality of the data included in the NIPS dataset was much better. Specifically, the NIPS proceedings included full text proceedings as opposed to only articles' abstracts that were collected for the Web of Science datasets. For the purposes of making comparison more straightforward, we depict the results from the full text k-means clustering for the NIPS proceedings below.

Table 11 lists the top words and the top proceedings titles for the first cluster identified[96] in the k-means clustering for the NIPS proceedings dataset. Identifying clusters in this dataset was perceived to be more difficult than for any of the abstract datasets scraped from the Web of Science. The current cluster was identified to be most closely aligned with the topic of cognitive neuroscience. However, the NIPS conference is in essence a conference on computational

[95] Again, because this is a very intensive process and the focus of many rigorous studies independently, the six examples given are all that were considered within the scope of the study, and none of these was considered to be for purposes other than demonstration and exploratory analysis. For exploratory purposes, no further defining of clusters was conducted (*n* was equal to ten for both k-means and LDA).

[96] The first cluster identified in each of the examples detailed in this section is used. The reason for this is to ensure that unbiased data is presented, and cherry-picked results are not shown to over inflate the quality of the demonstration.

neuroscience[97], so this cluster is not very informative regarding the nature of the dataset. Moreover, a second cluster on computational neuroscience was also identified, and when projected down to two dimensions this cluster was spatially adjacent to this first computational neuroscience cluster.

Table 11: Top Works and Titles for k-means on NIPS Dataset

Top Words	Top Proceedings' Titles
neuron	Ocular Dominance and Patterned Lateral Connections in a Self-Organizing Model of the Primary Visual Cortex
spike	Correlated Neuronal Response: Time Scales and Mechanisms
network	Onset-based Sound Segmentation
model	Neuron-MOS Temporal Winner Search Hardware for Fully-Parallel Data Processing
input	When is an Integrate-and-fire Neuron like a Poisson Neuron?

t-SNE was also utilized for visualizing the results of this k-means cluster analysis. Figure 17 depicts the clustering results projected down to two dimensions using t-SNE. The ten clusters can be seen, identified in the legend, including the two ‘computational neuroscience’ clusters which are spatially adjacent to each other. Spatially adjacent to both of these clusters is the ‘neural networks’ cluster, as would be expected. Between this region and regions of strong mixing representing “algorithms?” and “models?” there is a very heterogeneously mixed region. The question marks were included in the labels to indicate severe uncertainty regarding the label. “Reinforcement learning,” “clustering” and “kernel methods” are relatively isolated from the other clusters, with numerical algorithms also appearing isolated but to a lesser degree.

[97] In recent years, with the rise in deep learning, NIPS has become a home to any neural network related research.

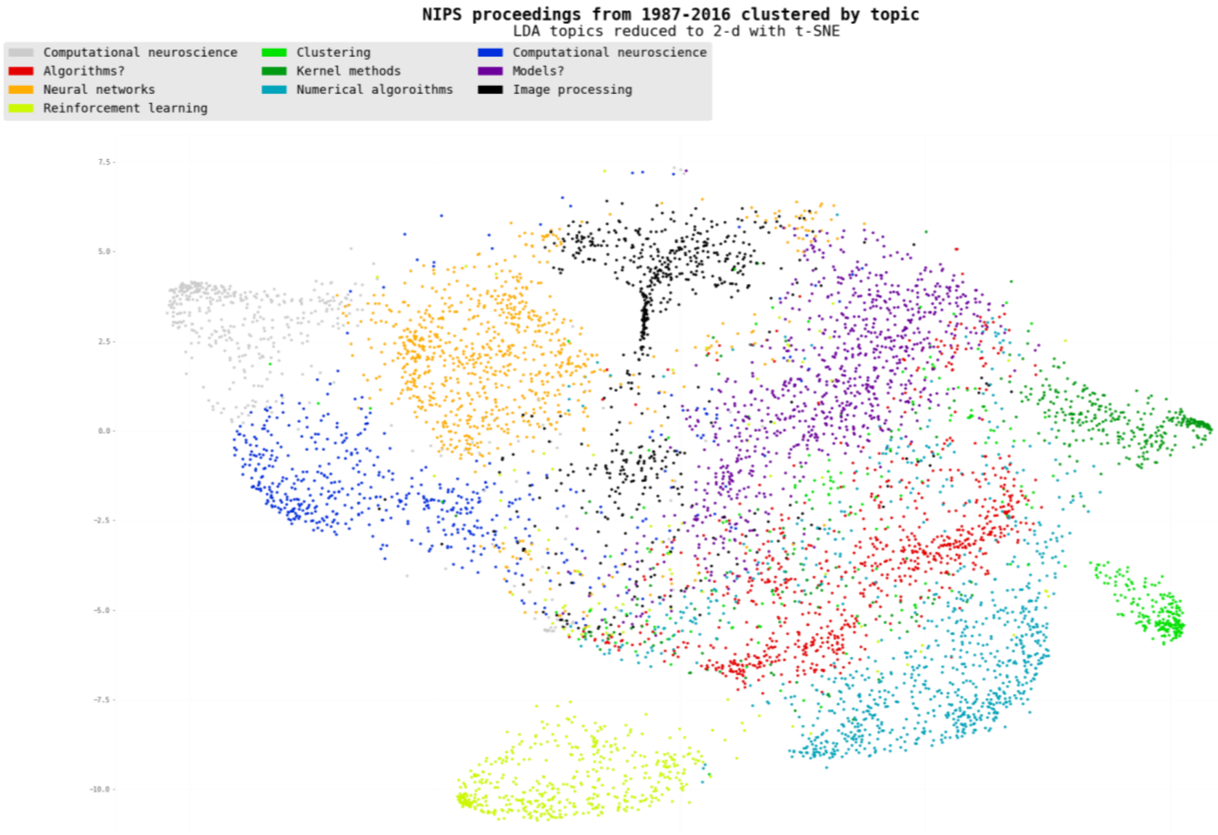


Figure 17: t-SNE clustering for the NIPS proceedings dataset.

Figure 18 depicts the evolution of the clusters from this k-means clustering analysis over time. Because the data includes thirty years of NIPS proceedings, the results are depicted at 5-year intervals. “Neural network” and “computational neuroscience” research appear to have been the primary foci in the early years of the conference, but more diversity has been observed in the more recent years. “Image processing” research has certainly increased in the past five years, but it is difficult to tell the degree of this increase because of an anomaly in the results. “Kernel methods” really took off between 2001 and 2006, which is consistent with the research interest in support vector machines at the turn of the century. Most recently, the “algorithms” and “numerical algorithms” topics have seen the most significant increases. This is also consistent with the algorithmic progress necessary to adapt neural networks to the highly parallelized implementations of backpropagation and optimization which are drivers for the deep learning research boom.

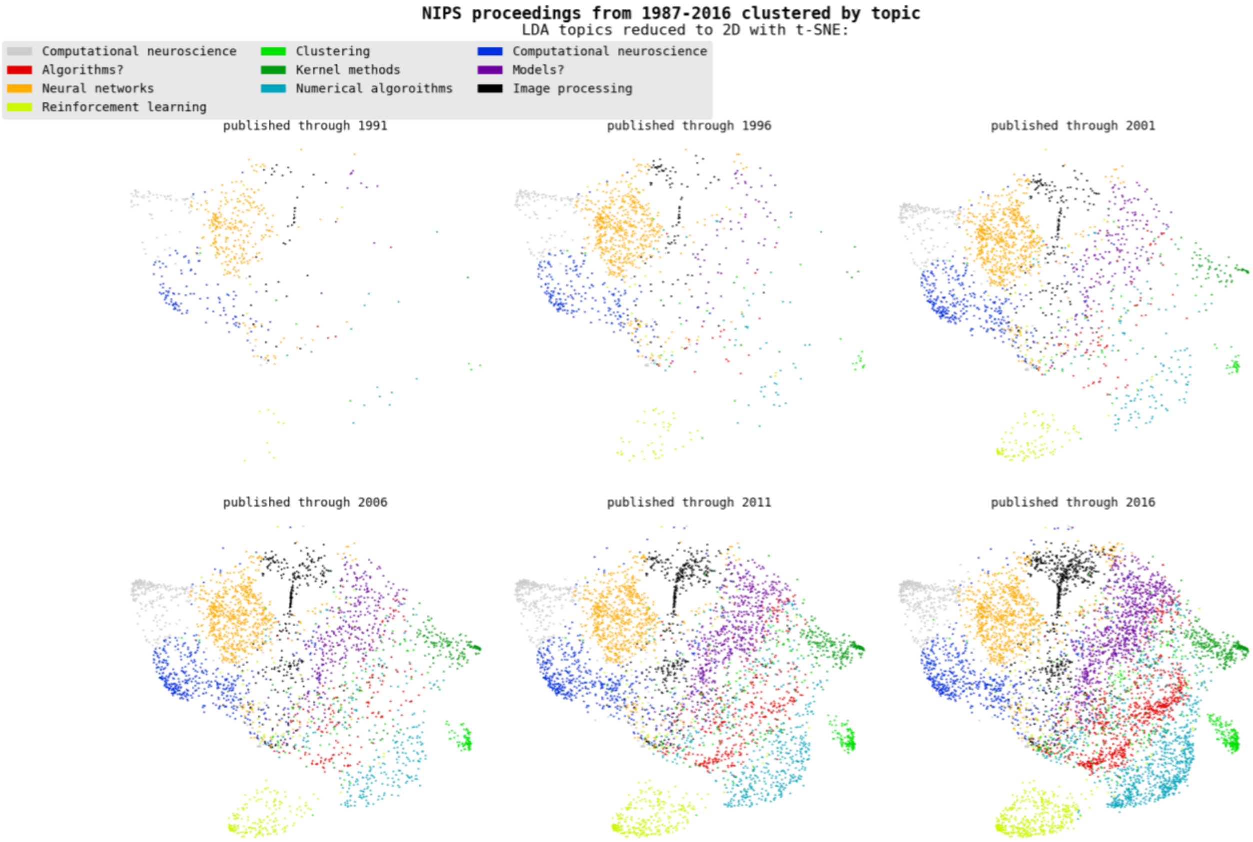


Figure 18: A t-SNE visualization of the development of the k-means NIPS proceedings clusters over time, at five-year intervals.

6.3 Extrapolation

The indicator selected for demonstration of this technique was Google DeepMind's (the world's foremost AGI research laboratory) operating costs. Records for DeepMind's data are available publicly because it remains a wholly owned subsidiary of Google based in London. Legally, Google DeepMind is required to report all of its basic financial data to the UK government through the Companies House each year. DeepMind was founded in 2010 by Demis Hassabis, Shane Legg and Mustafa Suleyman. It was acquired by Google in 2014 for reportedly over $500 million (Shu

2014). Figure 19 depicts Google DeepMind's operating costs in red, with the extrapolated trendline shown as a dashed black line. The y-axis is shown on a log scale.

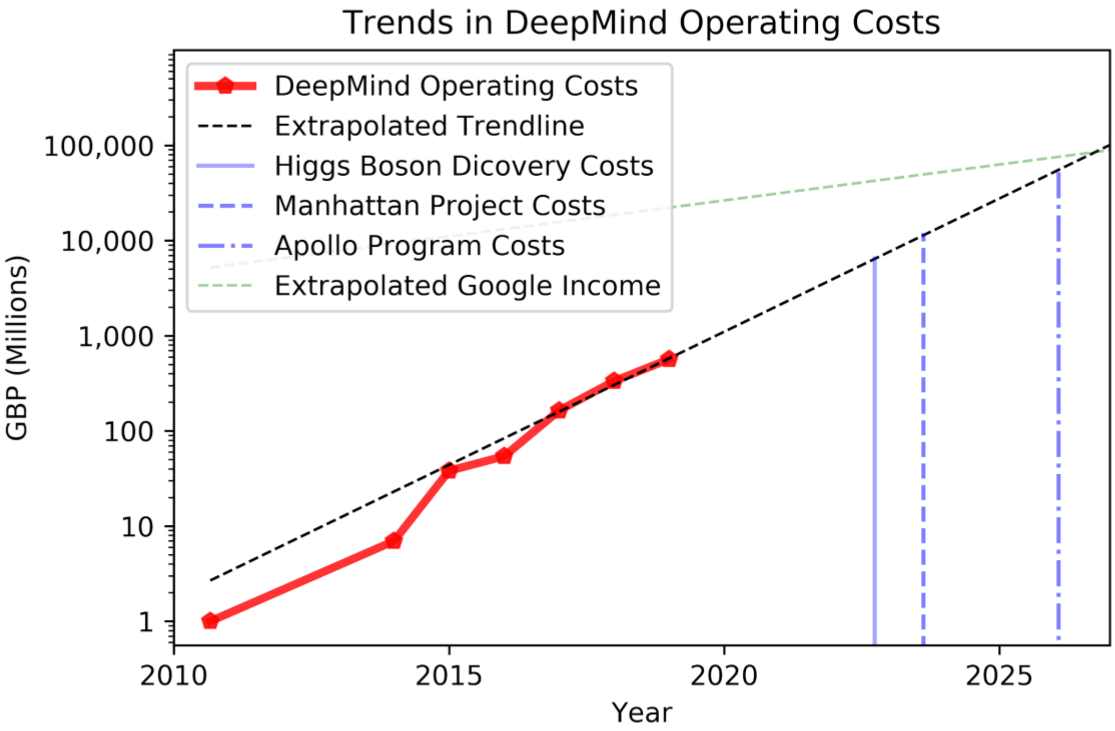


Figure 19: DeepMind's operating costs through 2019 extrapolated. Different large-scale scientific research projects cumulative costs are marked adjusted to 2019 British Pounds. The y-axis is depicted on a log scale.

The plot in Figure 19 reports operating cost in British pounds versus the year. It also depicts major scientific research projects for which there is no comparison such as the construction of the Large Hadron Collider in search of the Higgs Boson, the Manhattan Project and the Apollo Program (each depicted in blue and blue dashed lines, from left to right, respectively). Google's projected income is also included in the plot, shown by the dashed green line. Because Google does not pay dividends, this suggests that Google can afford to support a DeepMind research project on the scale of the historic research projects depicted by the blue dashed lines if they choose. Figure 20, below, depicts the projected DeepMind Revenue, and suggests that an 18-month hiring freeze could pare the discrepancy between revenue and operating costs. Moreover, there is reason to suspect DeepMind revenue is not representative of DeepMind's value to Google (Gruetzemacher 2019b).

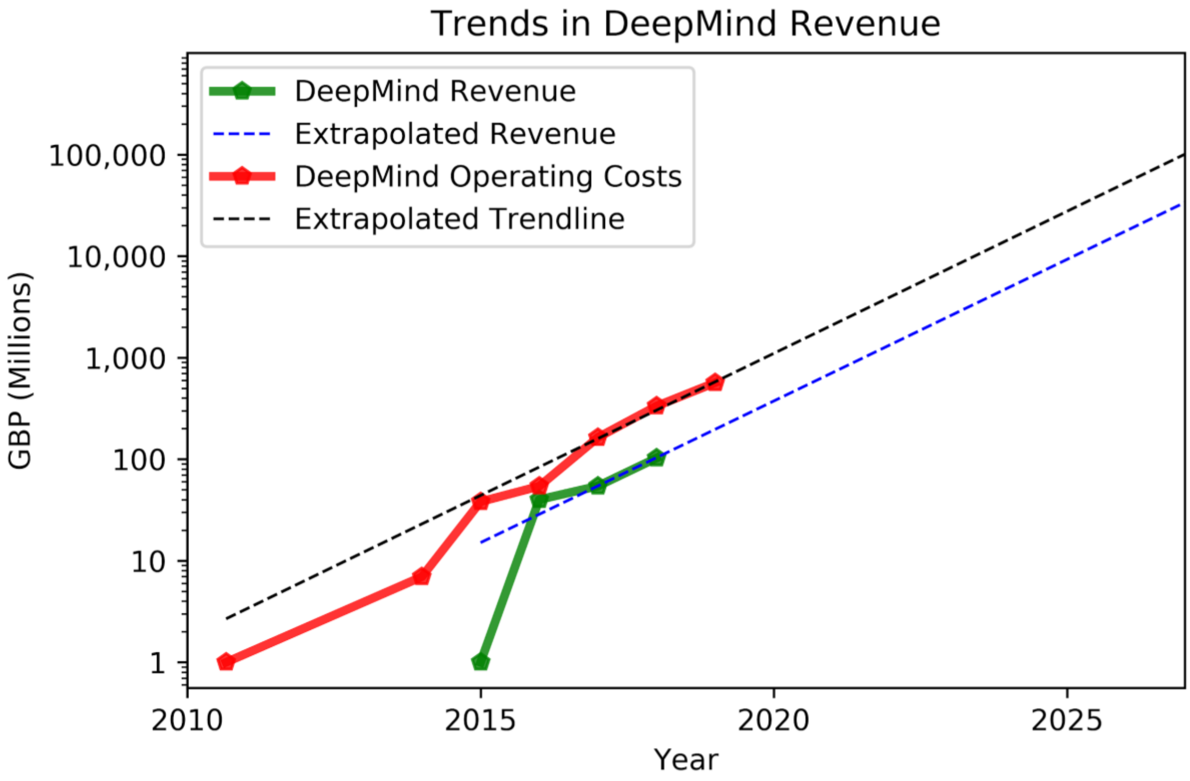


Figure 20: A comparison of projected DeepMind revenue and operating costs. The y-axis is depicted on a log scale.

6.4 Mixed Methods

The mixed methods portion of the study involved 20 interviews that were conducted using the same questions as in the survey as well as two additional qualitative questions. The interviews were conducted at the two conference locations where participants were solicited for participation in the survey, as well as at the 2018 NeurIPS conference in Montreal, Quebec, Canada in early December of 2018. Eight participants were interviewed at the Stockholm conferences, four were interviewed at the Prague conferences and eight were interviewed at the Montreal conference. The majority of one interview from the Prague conferences was lost due to failure of the recorder device. Furthermore, two of the interviewees did not feel comfortable making any probabilistic forecasts. Thus, much of the data was not suitable for quantitative analysis, but quantitative analysis was conducted for the data that was suitable for it (this did not include the forecasts). This is described in the first section. The second section then details some specific interviews and the third section details the qualitative questions.

6.4.1 Quantitative and Statistical Analysis

Data on the length of response time for each of the quantitative survey forecasting questions was available for all participants but the one who is recording was lost due to a technical issue. After reviewing the discussions during these questions, one participant appeared to have answered very quickly and spent a significant amount of time discussing material that was not relevant to the process of making their forecasts. Participants who did not give probabilistic forecasts still replied to each of the forecasting questions, giving a qualitative answer. The time for each interviewee's response to the survey questions[98] was then used for quantitative analysis.

Quantitative analysis was not conducted for the different forecasts because the sample size was too small for comparison when considering the lost interview and the two interviewees who were not comfortable making probabilistic forecasts. Moreover, four of the remaining interviewees found there to be no difference between forecasts of narrow AI systems and broadly capable AI systems. Three even anticipated that 99% of tasks would not be possible with narrow systems and would only be possible for a broadly capable system. Others pointed out that the ability to combine a large number of narrow systems to function as a single system may be relatively simple compared to creating a general intelligence. This was an interesting observation that underscores the importance of selecting the appropriate terminology and definitions for use in expert elicitations,

[98] This was computed by specifically tracking time spent on survey questions and not including time spent on the other interview questions, the introduction period or the debrief period.

and the importance of specifically considering TAI/DTAI/RTAI in the context of mid- to long-range AI forecasting.

In order to compare the amount of time spent by survey participants and the amount of time spent by interviewees on the same questions, it was necessary to return to the survey data to extract data on the time required for completion. The time for completion of the survey had a very wide range, from just several minutes to over a week. Obviously, participants who completed the survey online were much more prone to distraction. This made analysis challenging because assumptions were necessary in order to exclude a large portion of responses that likely involved the respondents being distracted or pausing to do other things before returning to complete the survey. Ultimately two different techniques were selected for resolving this issue; initially the data was viewed as a histogram of the length of time spent on each survey for all surveys completed in less than three hours. This is depicted in Figure 21, with bins for each hour and the first hour split into separate bins for survey completion times shorter and longer than the longest survey response time recorded during an interview.

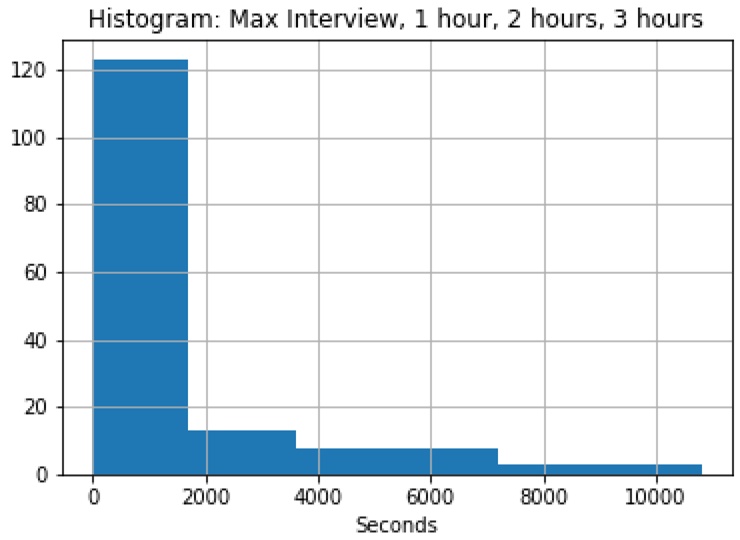


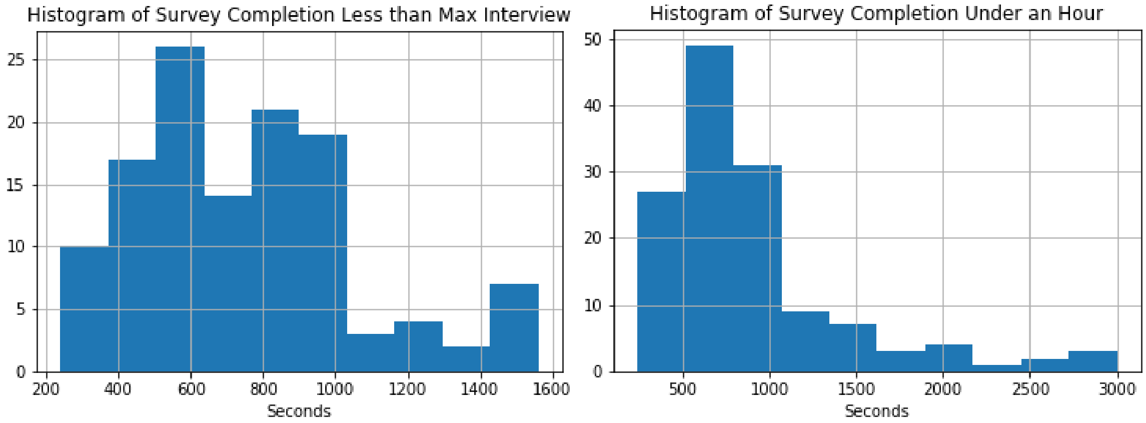
Figure 21: A histogram including four bins. The first bin is for survey completion times less than the time of completion for the longest interview. The second bin contains responses greater than the max interview length but less than an hour. The second and third bins are for one-to-two hours and two-to-three hours.

It was not possible to get a usable estimate of the portion of participants who likely became distracted. Nor was it possible to determine an estimate for the mean amount of time a participant may have been distracted. Thus, two assumptions are proposed:

1. Only a small number of participants spent longer actually completing the online survey without being distracted than the interviewee who took the greatest amount of time responding to the questions.
2. No participant took over an hour to complete the survey without being distracted, thus any responses of this length were not completed at a single time[99].

Based on these two assumptions it might be expected that the true maximum time for completing the survey likely fell somewhere between the maximum interview completion time and one hour. However, there is a significant amount of uncertainty here, and it is not unreasonable to postulate that no survey participant spent more time than the maximum interview completion time. If attention on digital devices was assumed to generally be poorer than during an in person interview, this could be a reasonable assumption, however, this analysis will only require the two assumptions listed above. Figure 22 depicts histograms for responses less than the maximum interview completion time (22a) and for responses completed in less than one hour (22b).

[99] It is useful to note that each link sent to potential participants was unique. So, if participants clicked the link to begin the survey, then closed the window and returned through the link at a later time, their progress was saved but the elapsed time was still computed from when the survey was started (i.e. it did include the time the window was closed).



(a) (b)

Figure 22: a) Histogram for survey completion times less than the maximum time to complete survey questions during an interview and b) a histogram of the survey completion times less than one hour.

In order to test whether there was a difference in the time taken by interviewees to complete the survey questions and the time take by survey participants to complete the questions, a two sample t-test for equivalence was conducted for each of these samples: survey completion times less than the maximum interview completion time and survey completion times less than one hour. A total of 123 respondents completed the survey in less than the maximum time to complete the survey during an interview, and 136 in less than an hour.

The p-values for the survey completion times less than the maximum interview time and those completed in less than an hour were found to be 0.000 and 0.024, respectively. Thus, in both cases there was strong statistical evidence to suggest that the survey respondents completed the survey questions in a much shorter amount of time. The mean for the survey completion times less than the maximum interview completion time, survey completion times less than an hour and the interviewee completion times were 12.5 minutes, 15 minutes and 21.4 minutes. Moreover, the average completion time reported by SurveyMonkey was 13 minutes[100], which would suggest that the assumptions used for this analysis were valid.

6.4.2 Qualitative Analysis

There were several interesting comments from participants during interviews regarding the forecasts for extreme labor displacement scenarios. Two participants asked if they could update their previous forecasts after being asked questions about more extreme labor displacement scenarios[101], while four further participants, of the nine who completed usable probabilistic forecasts, either requested or used their responses for previous levels of extreme labor displacement to anchor their responses to more extreme labor displacement scenarios[102]. Some participants simply replied to the question by indicating that some number of years be added to one of the previous forecasts.

6.4.2.1 Qualitative Questions

The qualitative questions asked each participant about anticipated progress and milestones in the next 5 to 10 years, as well as progress and milestones leading up to the development of the most extreme labor-displacing AI scenarios. These questions elicited a wide variety of responses. Some of the more interesting results for each of the three questions are discussed briefly below.

Regarding the first question on the most interesting milestones or indicators of progress reached over the past two years, some of the most interesting responses suggested that the best

[100] Assuming that SurveyMonkey's internal completion time computation is correctly accounting for inactive time spent on the page.

[101] One participant explicitly noted that they thought priming (Kahneman 2011) to be an issue in the questions. This was interesting because an initial plan for the survey involved randomly switching the order between the 90% forecasts and the 99% forecasts in order to evaluate the effects of priming. However, the decision was made to forgo this due to fears that it could be perceived negatively and lead to participants not completing the survey who otherwise would have.

[102] For example, when asked about 99% narrow systems participants would ask for a reminder of their forecasts for each quantile for 90% narrow systems.

indicators of progress were not always things reported in the news. For example, multiple participants suggested that games were not a good indicator of progress despite the significant news coverage that they had received. Another interviewee suggested that superhuman performance on games or other benchmarks was not necessarily as interesting as large jumps in performance on a benchmark to levels that were very close to human performance because superhuman performance on many games or benchmarks had recently become much more gradual than in the earlier part of the decade. Most interviewees responded to this question by describing some widely reported milestone or benchmark that was closely related to their specific research focus.

Regarding the second question on what participants anticipated being the most important signs of true progress over the next five years, again a wide variety of responses were recorded. Most interviewees were intrigued by the language used in the question, specifically by the inclusion of the word “true.” Initial responses to the question from many suggested that their answers may have been different had the word true not been included before progress, though it is doubtful this would have been the case. However, it does underscore a common perception that progress in AI as reported by the media or interpreted by those outside the AI research community was not an accurate reflection of how AI researchers perceived the progress. There may not be much signal from this common trend, but it was surprising to see the same reaction from over half of the participants being interviewed.

Some suggestions for true progress over the next five years did include continued progress on milestones and benchmarks, similar to the rapid progress of the preceding five years but respondents who suggested this might be the case did feel that the progress would continue at a constant rate rather than exhibiting discontinuities in the rates of progress. Others suggested that true progress would not be in the realms of vision or reinforcement learning, but rather in NLP and robotics. Others suggested that a sign could be the integration of reasoning techniques with some of the current leading machine learning techniques (e.g. embedding symbolic reasoning in deep learning). These last two categories mentioned now seem to have been prescient as, based on conversations from most of the final conferences attended as part of this study, these are three of the most highly discussed topics in the AI community.

Regarding the final question, on milestones anticipated to precede the most extreme levels of labor displacement, not all participants were confident providing answers. Other respondents openly discussed their uncertainty of whether there would be any canaries in the coalmine to signal the arrival of these powerful systems[103]. Others saw advanced AI technologies like metacognition and theory of mind to be indicators of approaching these extreme labor displacing scenarios. However, in general, responses included hesitation and appeared to be difficult to answer, even for the experts who were working directly on technologies with the potential to lead to extreme labor displacement.

6.4.3 Selected Interviews

In this section three selected interviews are highlighted. Each of the interviews highlighted here was selected because of the specific expertise of the participant as well as their calibration question performance and their forecasts. While some of the participants performed well, or poorly, on the calibration questions this should not be thought of as a good indicator of the value of their forecasts. Rather, since there were so few forecasts not much can be said about the calibration performance[104]: this is discussed later. It is also significant to note that there are many other

[103] For further reading on potential canaries for preceding RTAI please see Etzoni (2020).

[104] This is another reason why the naive calibration technique is better suited for attempts to include calibration in surveys/interviews.

interviews which would have been interesting to highlight, but, given the vast number of interviews conducted in this study, only these three were selected for demonstrating the variety of results from the variety of experts surveyed.

One particularly interesting interview was with one of the leading natural language processing researchers, well known for developing the first neural word vector model (Mikolov et al. 2013). This interviewee completed all of the evaluable calibration questions accurately and gave very left skewed timelines that were thirty and forty years past the median for all of the extreme labor displacement scenarios. While the interview was lengthy, his forecasts were all made quickly, simply building off his initial forecast for 50% labor-displacing AI. He focused on how current progress was not indicative of true progress in AI. He strongly believed that progress on games and other benchmarks widely thought to be indicators of strong progress in AI were misleading and were toy problems. Speaking of projects at OpenAI and DeepMind he stated "it's actually not aiming at convincing researchers like me, it's aiming at convincing people there that something amazing is happening." One of the most interesting aspects of this interview was the fact that when asked specifically about progress in NLP, his most prominent area of expertise, he expressed that he thought good results could be generated with large amounts of money[105], but that these results would not be indicative of true progress toward natural language understanding.

Another interesting interview involved a new research scientist at a leading AGI research laboratory. This interviewee was poorly calibrated, based on the evaluable calibration questions, getting each incorrect, and had forecasts that were right skewed and involving very short timelines relative to other responses. Specifically, striking forecasts were three- and four-year forecasts for 10% probability of 90% and 99% broadly capable systems. Moreover, this participant did not feel

[105] This seems incredibly prescient given the recent progress demonstrated by OpenAI with GPT-3 (Brown et al. 2020).

that 99% of tasks could be accomplished without a broadly capable system, or, without generality. This participant, working directly on these technologies, expressed that he "wouldn't necessarily expect to get so much advanced warning" about the arrival of DTAI or RTAI.

Yet another interesting interview involved a robotics expert who was a pessimist regarding quantitative forecasting. He found it difficult to ascribe probabilities to future forecasts but was still able to make qualitative forecasts. One particularly interesting qualitative insight of his was mentioned in regard to the two calibration questions which focused on self-driving automobiles[106]. He stated that edge cases would be problematic, and that this would greatly delay the viability of self-driving vehicles. While he did not give point estimates for each of the self-driving vehicle calibration questions, his discussion of these questions focused on whether or not the probability would be non-zero, suggesting a very strong confidence in the significance that edge cases would play in future development of self-driving vehicles. Since this interview, the challenges posed by edge cases have become widely discussed in the machine learning research community and have become generally accepted to be a problem for the deployment of self-driving vehicles that will delay their widespread use by a decade[107].

6.5 Scenario Network Mapping

The SNM demonstration involved conducting three separate SNM workshops in order to explore the possibilities of different variations of the SNM workshopping process. The locations and dates of these workshops have already been noted in section 4.6. Here, we report the results from each

[106] In the time since these interviews, it has come to be more widely regarded that self-driving vehicles pose more of a challenge than previously thought. Specifically, many now believe that solving the first 95% of cases was straightforward while solving the final 5% becomes increasingly more difficult. In fact, at the International Conference on Robotics and Automation in 2019, a keynote speaker from Uber's self-driving vehicle laboratory encouraged interested students to begin their PhDs in self driving vehicle research because so much work remains (particularly on edge cases). However, at the time of this interview, this was an opinion only known to roboticists closely working on the problem.

[107] Another interviewee suggested that he anticipated self-driving vehicles to become widely available sooner, but that they would not be able to drive in all road conditions without remote assistance for over a decade.

of these workshops as well as some of the lessons learned from this experience. Although each of the workshops began with a specific outline for the intended workshop process to be implemented (see Appendix E), modifications were made for all cases to adapt to numerous unanticipated circumstances based on the participants, event scheduling and the event workspaces. All participants were volunteers who were being weakly rewarded for their time, so there were no strong incentives to ensure continued participation through all sessions or to complete more or longer sessions in accordance with suggestions by List (2005).

There was a great amount of diversity among the different groups of participants for the developmental SNM workshops. Table 12 depicts a rough breakdown of the participants' different backgrounds for some context before beginning the reporting of the results. The distinct makeup of the different groups is significant for considering the significant variance in the results of the workshops.

Table 12: SNM Developmental Workshop Information

Workshop	#1	#2	#3
Location	Avila, Spain	Marburg, Germany	Moscow, Russia
No. Participants	5 (7 initially)	11	12 (5-9 daily)
Expertise (proportionally)	AI student, AGI researcher	Neuroscience expert (7), ML Researcher (2), AI Strategy Researcher (2)	Transhumanist/NeyroKod Developer (7), AGI researcher (3), ML researcher (2)

6.5.1 SNM Workshop I

The first of the three SNM workshops was the one that was the most closely related to the specific workshopping process that was described by List (2006). This involved the use of all physical materials for creating the map (e.g., post-it notes, string[108], etc.), and the raw results can be seen in Figure 23. The workshop was held in conjunction with another event that allowed for two separate three hour workshops one day apart. Due to this, the workshop utilized a significantly modified version of the original outline[109] that had been adapted from List (2005). In the Figure there appears to be two different scenario network maps because the adapted version of the map involved focused initially on identifying significant progress in the recent past for two purposes: 1) to identify the most significant vectors of current progress and 2) to familiarize participants with the process of scenario tree development.

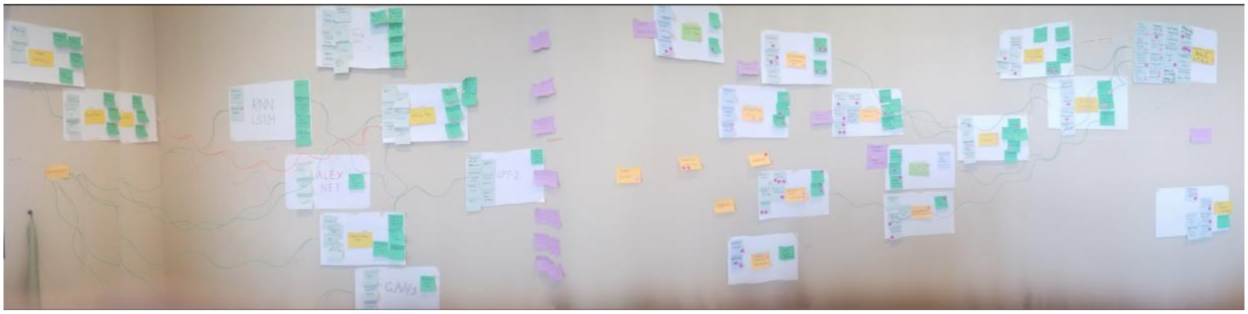


Figure 23: The physical scenario network map created from the first SNM workshop in April of 2018. This was the only scenario network map from the three workshops that was taped to the wall. The left side of this map tracked past technologies to the present, where the purple (i.e. present) technologies are lined up, just left of center. On the right side the orange post-it notes represent the next generation technologies identified by the larger group and used by a subgroup during the event tree creation period. The two yellow post-it notes at the far right side represent AGI (top) and brain-computer interface (bottom); these each have no leaves in the event tree and are end states for the map. The three yellow post-it notes to the left of these end states were technologies that were identified by the backcasting subgroup. The weed trimmer line (blue and green, but difficult to see) connects the different event trees as appropriate.

This workshop was split into the two separate workshops by initially focusing the first workshop on how the past technologies have led to the future technologies, and then focusing on what the anticipated next generation technologies are and what the end states of the scenario network map would be. The second workshop focused on taking the next generation technologies

[108] The workshop was conducted in Avila, Spain and no string was available. Colored weed trimmer line was used in lieu of string because it was the best alternative that was easily accessible.

[109] The original outline can be seen in Appendix E1. The actual process combined the second and third sessions depicted in Figure E1.1.

and generating event trees by midcasting and backcasting. Following the generation of event trees the group determined which event trees should be connected, and the weed trimmer line was taped to the different event trees accordingly.

The creation of event trees is the most essential step to the SNM workshopping process. The results depicted in Figure 23 are difficult to see in detail, thus Figure 24 is included so that the resulting event trees can be seen more closely. The event tree pictured is for meta-cognition, or, in other words, thinking about thinking. The roots are shown on the left on the light blue post-it notes. The leaves are shown on the right on the green post-it notes. The trunk of the tree, here, meta-cognition, is on the wider post-it in the center (this post-it is yellow because this was a backcast technology).

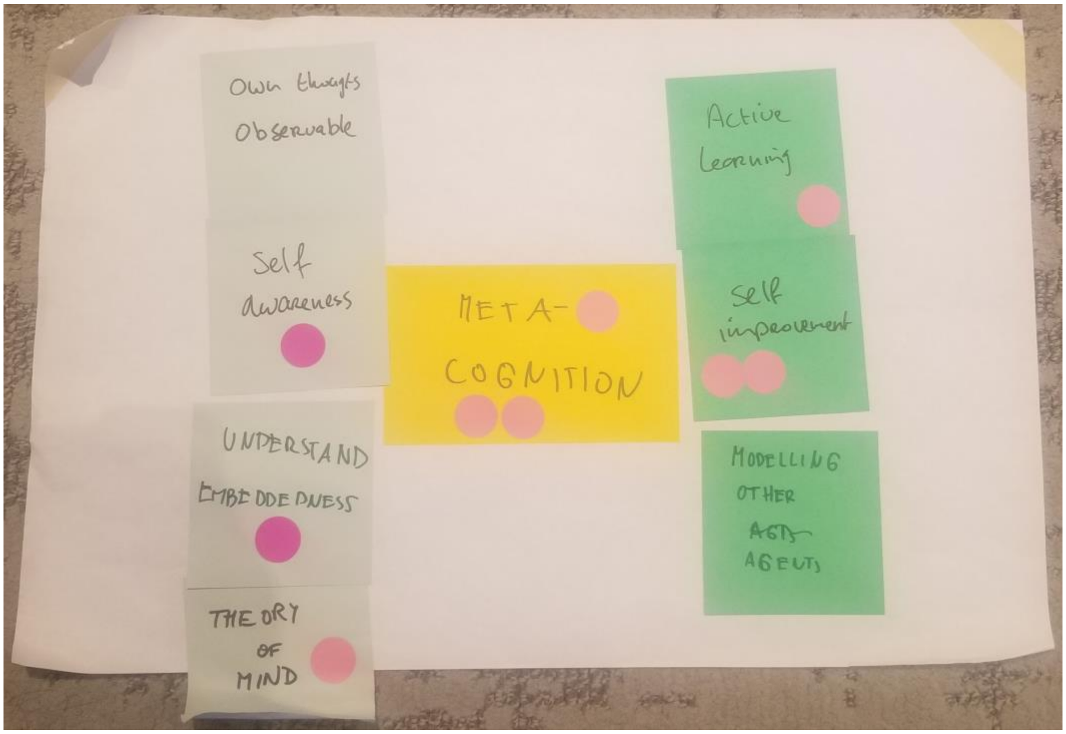


Figure 24: This is an event tree created during the first SNM workshop for the technical capability or technology of meta-cognition. The roots of the tree are shown on the left on blue post-its and the leaves are shown on the right on green post-its. Meta-cognition is the trunk of the tree and is depicted in the center on the yellow post-it.

On the post-it notes in the event tree depicted in Figure 24 circular stickers can be seen that were used for voting on the most significant topics. Each participant had 10 stickers of each color to use in the voting. The light pink stickers were used to represent which technologies the participants thought would be the most important for inclusion as trunks of event trees in the overall scenario network map. Because of the shortened format of the workshop given the constraints previously described, a follow-up round for pruning, reorganizing and adding event trees was not possible. However, if time had not been a constraint this would have been desirable. These stickers and the ability for participants to vote on technologies identified in the event tree creation process for further decomposition was included to be consistent with List's focus of the holonic principle in the SNM technique. The root in Figure 24 that contains a pink sticker was a trunk in another tree in the scenario network map that the participant may not have seen before placing the sticker. Participants were not instructed to review the existing event trees prior to voting, which may have been helpful to the facilitator and was heeded in the 2nd workshop.

The darker pink stickers were allocated to participants during voting so they were able to identify things that they thought were significant components of the map, but that they did not feel were important enough for their own event trees. Two of these darker pink stickers are shown in the event tree in Figure 24 on the previous page, while substantially more are shown in Figure 25 below. Figure 25 is included here to illustrate the end state event tree with only roots (here shown for AGI[110]). The figure also illustrates more clearly than in Figure 24 how the weed trimmer line and masking tape was used to connect the different trees in the map. While blue and green trimmer lines as well as masking tape were used to depict the connections between the different trees, there is not any implicit difference between the different types of connections. The masking tape was

[110] Due to time constraints the other end state identified was not included in the map.

useful in that it enabled storing more information about the connections (which would have otherwise been lost) because it could be written on. However, that information was present for all of the other connections and was simply not stored[111]. Both blue and green trimmer lines were used in order to speed up the process of connecting the different event trees. The process was more time consuming than expected, however, all participants were enthusiastic about continuing the workshop session past the allotted time in order to complete the map[112], and they all took initiative to actively contribute (this was a departure from suggestions by List that the facilitator or co-facilitator would be the only person to actively connect the different event trees, only with consensus from the group).

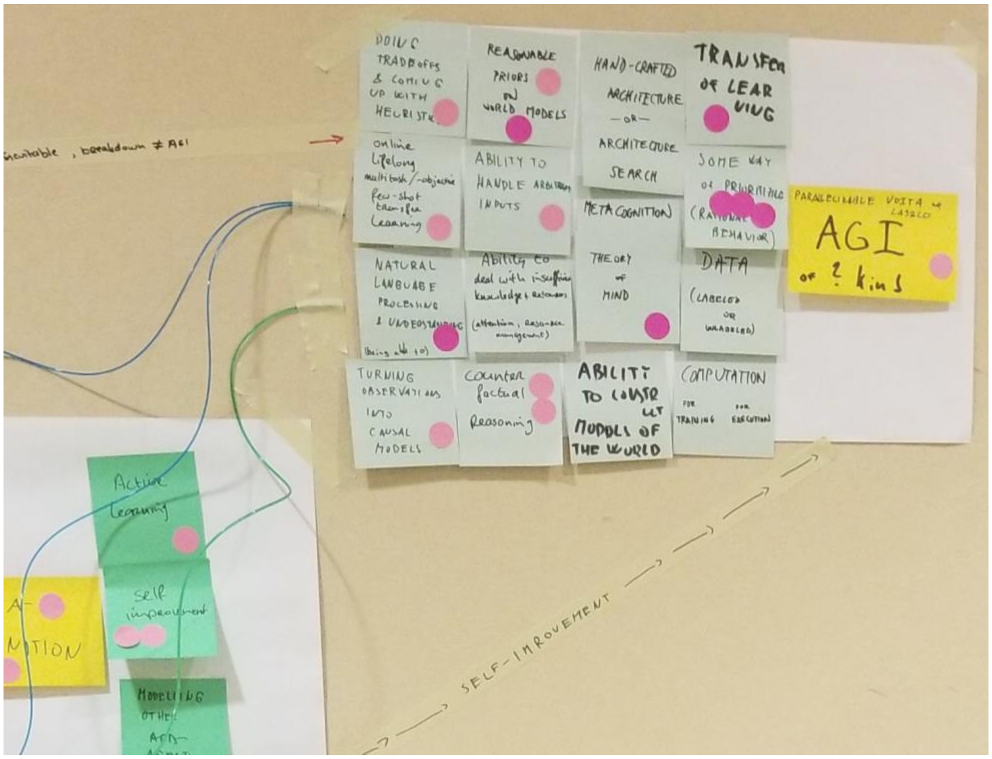

scalable , breakdown ≠ AGI
DOING TRADEOFFS & COMING UP WITH HEURISTICS
REASONABLE PRIORS ON WORLD MODELS
HAND-CRAFTED ARCHITECTURE —OR— ARCHITECTURE SEARCH
TRANSFER OF LEARNING
online lifelong multitask/-objective few-shot transfer learning
ABILITY TO HANDLE ARBITRARY INPUTS
META COGNITION
SOME WAY OF PRIORITIZING (RATIONAL BEHAVIOR)
PARALLELIZABLE
AGI
NATURAL LANGUAGE PROCESSING & UNDERSTANDING (being able to)
Ability to deal with insufficient knowledge & resources (attention, resource management)
THEORY OF MIND
DATA (LABELED OR UNLABELED)
TURNING OBSERVATIONS INTO CAUSAL MODELS
Counter factual Reasoning
ABILITY TO CONSTRUCT MODELS OF THE WORLD
COMPUTATION FOR TRAINING FOR EXECUTION
Active Learning
Self improvement
MODELLING OTHER
SELF-IMPROVEMENT

Figure 25: This image shows the corner of the event tree detailed in the previous Figure as well as the focal end state for this workshop. Five different event trees converge on this end state as can be seen from the weed trimmer line and the masking tape directed into this node of the scenario network map.

[111] Due to the large number of connections, and the tangled web that results, despite its benefits masking tape would not be practical for use with all of the connections.

[112] This alone is a good indicator of the success and perceived utility of the participants regarding the overall workshopping process.

The group who completed this workshop was small, but each participant had substantial technical expertise, more so than the participants in either of the remaining SNM workshops (see Table 12). Moreover, these participants were all actively involved in research on AGI safety, so they were all familiar with at least some portion of the broad body of work related to different hypothetical techniques for developing AGI. Consequently, the results were the most technical of the three workshops and were thus the best example of how the modified method can be used[113] to map plausibly future technical progress of a technology.

While the NeyroKod software was in lieu of the physical materials for the third workshop, the results from each SNM workshop were all digitized using this package that was described in section 4.7. The results of this digitization can be seen in Figure 26. Eight of the event trees were connected during the end of the 2nd workshop session, however, some were left out because the workshop session was already over its projected time and there was still the need to debrief.

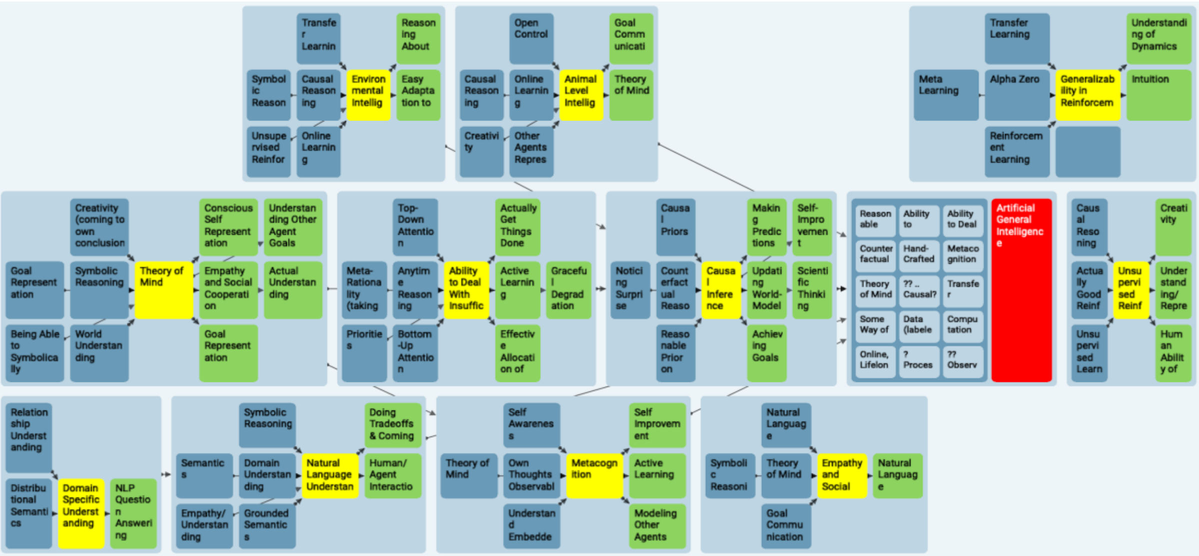


Figure 26: The digitized version of the scenario network map generated from the 1st SNM workshop. The end state, AGI, is the red block in the event tree with only roots. Three of the event trees were not connected in the map during the workshop due to time constraints. The unconnected event trees are all pushed to the right because the software is still in Beta development and all instructions are in Russian. Despite these limitations, the strength and suitability of NeyroKod for the digitization of the results from SNM workshops is clear.

[113] As suggested by List (2005), but previously undemonstrated.

A debriefing was conducted with participants after the end of the 2nd workshop session during which participants were questioned regarding their experience and perceptions about the workshopping process and the results of the process. Participants' perceptions were generally positive, with some comments about elements that were less than ideal or that could be improved in future iterations of the workshop process. Four of five of the participants indicated an interest in participating again with more experts and for a longer period of time once the process had been refined further. The most significant thing recommended for improvement that had consensus among the group was the lack of utility of the initial exploration/mapping of historical technologies to the current technologies. Participants were all generally pleased with the insights that were gained from the full workshop process as well as the final result.

Participants were asked to make notes on their experiences. Only one of the five participants who completed the workshop complied with this, but this response came from the most experienced participant and was relatively detailed. The participant found the process particularly enlightening, because prior to the workshop he had perceived himself having a comprehensive understanding of the different paths through which AGI could be developed, but during the workshop he recognized many things that he had not previously considered because of the diversity of backgrounds of the other participants in the workshop[114]. Moreover, this participant's notes are consistent with the notes from the debrief suggesting that all participants' experience of the workshop was similarly positive; each of the five participants found the

[114] This participant was the most knowledgeable participant regarding AGI to have participated in any of the workshops. Many persons at the 2nd workshop were very knowledgeable, but they had not worked directly on AGI research for an extended period of over five years as this participant had. Thus, this participants' detailed response is the strongest evidence of the utility of the technique for those who are actively working on the development of AGI. It also suggests that the inclusion of experts with interests in AGI, but from different perspectives, was a very valuable aspect of the process.

experience enlightening and insightful in some way. Specifically, the participants all realized after voting that at least one topic which they had not previously considered to be important to the possible paths to AGI should be explored in more detail through expansion by the creation of more event trees in a follow-up workshop session.

6.5.2 SNM Workshop II

The second SNM workshop was modified slightly following the first workshop to try to incorporate elements considering the safe development of RTAI (e.g., AGI). Several challenges arose during the immediate workshop presentation that had not been anticipated during the planning period. For one, the workshop was conducted in a difficult environment (the attic of an old German farmhouse) that was poorly suited for the objective, against the advice of List (2005). Consequently, it was difficult to effectively position the developing map on a wall during the workshop, so the map was placed on the floor. This made it difficult for all participants to see the map during discussions. This was further complicated by the fact that there were over twice as many participants in this workshop as were in the previous workshop where an equally sized, better suited room was still not too large for that smaller group. For these reasons, the process that was implemented is not relevant to the results of this workshop because the value in the results comes from its failure rather than its success. Moreover, this failure was not consequent of the workshopping process, rather, it was a consequence of sample bias as well as the other challenges just mentioned.

This workshop was conducted with two two-hour sessions. Originally, two half-day sessions had been planned, but the nature of the hosting event allowed for participants to vote on the activities of the research retreat and their scheduling. This had some further negative consequences on the results of this workshop. Particularly, it forced this workshop to be conducted in a single day, without time for reflection that the original instructions call for (List 2005). It also forced participants to try to create an entire map with only a thirty-minute lunch break in between the different sections. Each of these deviations from the normal approach to the workshop seemed to have a significant impact on the experience and flow of the workshop. Moreover, participants began the workshop with different levels of commitment and different levels of expectations for the results of the workshop because of the factors that were discussed in the previous paragraph. Because of this, completing the workshop in such a short period, without reflection time, may have led to lowered expectations from those who had previously been optimistic in the morning session. While the original plan had been to have two half-day sessions, the alternative plan was to have a morning session with a three-hour break, including a walk, before reconvening in the later part of the afternoon. This plan, however, was voted down by participants who were eager to finish, so the 2nd session after lunch was rushed, and the map was not completed for this reason.

Figure 27 depicts the results in a digitized form. However, the results from this workshop do not make a full map. The participants lost interest in the end of the 2nd workshop and chose to proceed with the debriefing rather than to continue over the allotted time to complete the map. Several distinct event trees as well as four unique end states are illustrated in the Figure. These elements reflect the attempts to incorporate paths to safe AGI development in the SNM process. The four unique end states are also the result of recommendations from the first workshop that different forms of AGI be defined for generating the map.

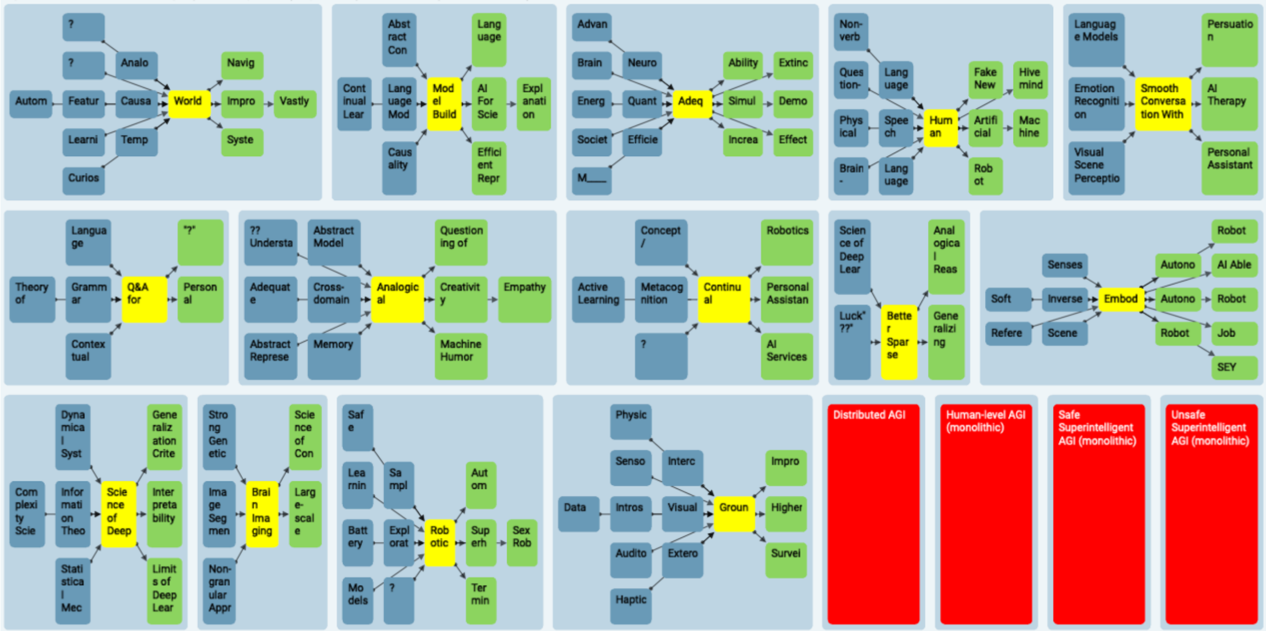


Figure 27: The event trees created during the 2nd SNM workshop. The four red blocks represent the four end states, each of which are different variations of AGI, some of which are safe and lead to outcomes beneficial to humanity while others are unsafe and lead to catastrophic or existential events.

The first SNM workshop relied solely on a debrief for assessing participants' experiences of the workshop and the outcome. For this second workshop, a survey was developed and administered to participants. Time was given during the debrief for participants to complete these surveys on their mobile devices, thus, there was a very high response rate of 10 out of 11 of the participants. Starkly contrasting the participants' experiences during the first workshop, participants of the second workshop were moderately to very dissatisfied with the experience and the outcome. There are too many reasons for this to mention specifically, however, there were some common themes. The first of these involves the fact that several participants believed the development of HLAI or AGI to be an impossible or impractical objective. The proportion of such participants was over twice as high as that of the sample population of survey participants attending the Stockholm conferences. This was a circumstance that was unanticipated and that created an adversarial element in the dynamic between the facilitator and the participants[115]. In normal proportions this may not pose a challenge, however, this is only conjecture because both of the other workshops involved only participants with firmly held beliefs about the development of AGI. This workshop was further complicated by the fact that the participants were allowed substantial flexibility with respect to the process of the workshop, including self-selecting groups – the skeptical participants all chose to be on the same team. Thus, one group was primarily focused on backcasting from different safe AGI outcomes while the midcasting group undertook a futile attempt to envision paths toward an objective they did not believe to be realistic or plausible.

Substantial discretion on the part of the facilitator was exercised in the first SNM workshop to shape the direction of the process because the participants' expertise was perceived to be novice. Participants' expertise for the 2nd workshop was perceived to be more substantial than it was because the facilitator had limited knowledge of neuroscience and its implications on AGI prior to the workshop. Thus, limited guidance regarding that perspective was possible, and, in order to evaluate the impact of this particular expertise on the resulting scenario network map participants were given significant freedom and flexibility. This was ultimately a failure, but it was effective at illuminating many different shortcomings of the technique and the impact of utilizing a sample poorly representing the broader AGI research community.

Another common comment was that the focus of the workshop itself was too technical and that more AI strategy relevant elements should have been considered. This was an interesting comment considering that the first SNM workshop had been purely technical and that this workshop had been adapted to consider elements of AI safety. However, challenges posed by collective action problems were not considered, and are something that SNM is suggested for in

[115] One participant had previously worked for a leading AGI research lab and had left due to theoretical differences with the organization.

the strategic planning portion of the holistic forecasting framework proposed here. These comments further suggest that these participants felt that the technique would also be effective for this purpose, which, given the general dissatisfaction of these participants, the fact that many still saw positive applications of the technique even during failure of its intended objective implies that they perceived the exercise to be more effective than other workshopping techniques they had experienced previously. Some participants expressed this more explicitly to different degrees.

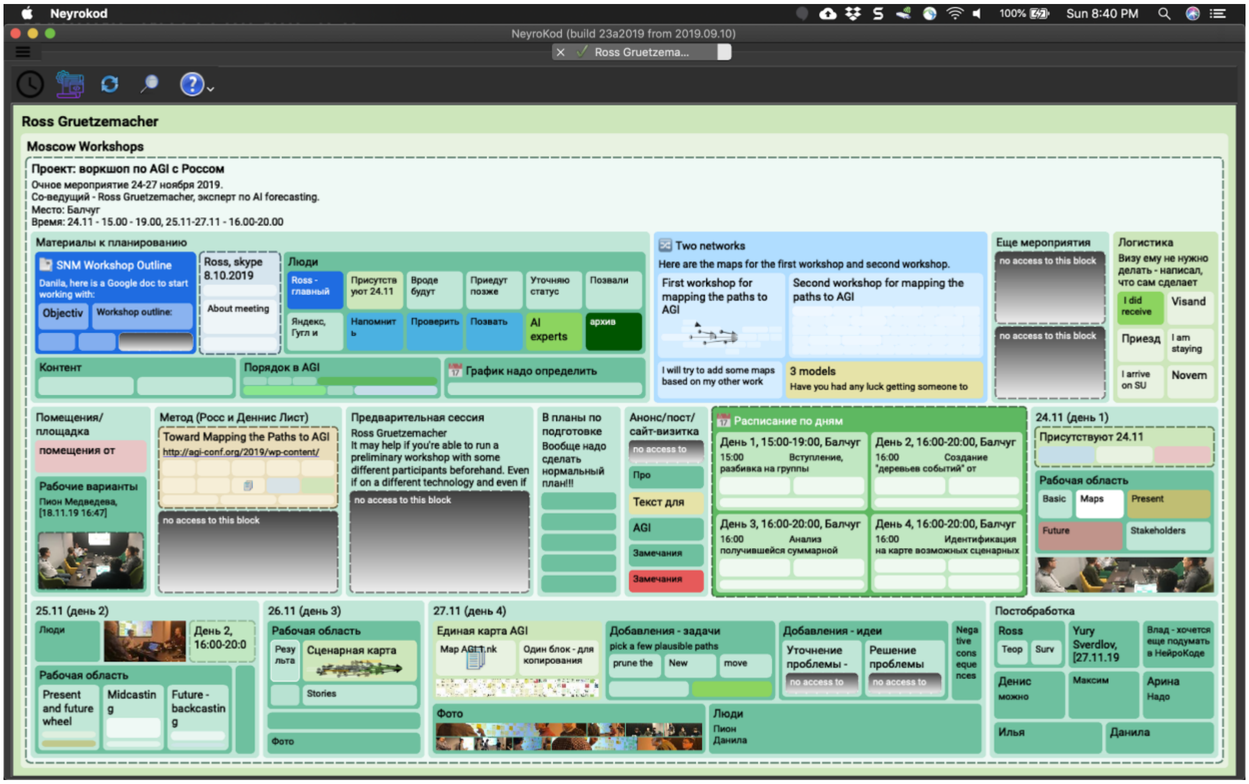
Of participants who were not entirely dissatisfied and who did not feel that the workshop was too technical, four did see potential for the process in the intended role of identifying plausible technical paths to AGI. While the workshop ultimately failed to generate a map, this was a positive development because it forced these participants to give effective criticism and suggestions for future improvement. The overwhelming lack of technical expertise in the area of machine learning was perceived to be a bottleneck from two, who felt that it could have been more productive if participants had been included with substantial expertise in this domain. Two other participants felt that it could be very effective for exploring paths to safe AGI, which was the purpose of the modification made in the early planning stage of preparation for this workshop. These, and other comments, could be very useful for any who attempt to conduct a SNM workshop with a more technical audience, or an AI safety-oriented audience, at some point in the future.

While this workshop was ultimately not successful, many useful things were learned that have value for further developing the technique. For example, it is incredibly important to have both the right space as well as snacks, so participants wo not try to rush in order to have a meal. Also, it is crucial to have experts that, despite their opinions, are open-minded and have positive expectations for the outcome of the workshop. Perhaps they may not feel that a map of the paths to AGI is possible, but they could be willing to contribute productively to others who want to create such a map and they could be willing to consider different applications of the technique that may be useful and that they think are possible (e.g., identifying paths to TAI/DTAI).

6.5.3 SNM Workshop III

The third SNM workshop yet again involved a highly modified process. Similar to previous workshops, the modifications were made on site based on the expertise of the participants and the resources available. Many of the modifications involved the inclusion of the NeyroKod software which is still in beta testing and can be used for mapping different futures. When used for the digitization of the results from the first SNM workshop, NeyroKod's value for this type of mapping was made very obvious. To be certain, digitization of the maps from the first two workshops was attempted with two other software products, also in beta testing; each of these were very poorly suited for the specific needs of the digitization of the scenario network maps generated during the workshops. List (2006) suggests several software packages that can be used for digitizing the results of SNM, but all of these were crude. The most successful attempt at digitizing the first map was made using Microsoft Excel, however, actually completing the map in this fashion would have taken an extremely long time and would have been difficult to perceive visually. Other, more visually acceptable techniques, e.g., Microsoft Visio, would have also involved an extremely long and tedious process, which, given the amount of data to process and analyze for this study, was very undesirable. NeyroKod solved all of these problems.

However, NeyroKod's strengths are not limited solely to digitizing maps that were created with tactile resources. NeyroKod is actually much more powerful than has been demonstrated in the digitization projects. It is a powerful collaboration tool that is unlike any software that the author has found thus far, and which has strong potential for numerous applications aside from mapping such as more general project management. Similar to Google Docs, it enables live collaboration on scenario maps. Because this workshop was conducted in collaboration with the NeyroKod developers, it was used for the purpose of project management for the third workshop. Figure 28 depicts the entire project management map. This map was shared over a network and the different sections were editable by those with the proper privileges. Editing of different components of the map were updated in real time, with very little lag for a beta product. Consequently, the full power of this novel software was able to be leveraged for this third workshop.



Ross Gruetzemacher

Moscow Workshops

Проект: воркшоп по AGI с Россом
Очное мероприятие 24-27 ноября 2019.
Со-ведущий - Ross Gruetzemacher, эксперт по AI forecasting.
Место: Балчуг
Время: 24.11 - 15.00 - 19.00, 25.11-27.11 - 16.00-20.00

Материалы к планированию

SNM Workshop Outline
Danila, here is a Google doc to start working with:
Objectiv
Workshop outline:

Ross, skype 8.10.2019
About meeting

Люди
Ross - главный
Присутствуют 24.11
Вроде будут
Приедут позже
Уточняю статус
Позвали
Яндекс, Гугл и
Напомнить
Проверить
Позвать
AI experts
архив

Контент

Порядок в AGI

График надо определить

Two networks
Here are the maps for the first workshop and second workshop.
First workshop for mapping the paths to AGI
Second workshop for mapping the paths to AGI
I will try to add some maps based on my other work
3 models
Have you had any luck getting someone to

Еще мероприятия
no access to this block
no access to this block

Логистика
Визу ему не нужно делать - написал, что сам сделает
I did receive
Visand
Приезд
I am staying
I arrive on SU
Novem

Помещения/площадка
помещения от
Рабочие варианты
Пион Медведева, [18.11.19 16:47]

Метод (Росс и Деннис Лист)
Toward Mapping the Paths to AGI
http://agi-conf.org/2019/wp-content/
no access to this block

Предварительная сессия
Ross Gruetzemacher
It may help if you're able to run a preliminary workshop with some different participants beforehand. Even if on a different technology and even if
no access to this block

В планы по подготовке
Вообще надо сделать нормальный план!!!

Анонс/пост/сайт-визитка
no access to
Про
Текст для
AGI
Замечания
Замечания

Расписание по дням
День 1, 15:00-19:00, Балчуг
15:00 Вступление, разбивка на группы
День 2, 16:00-20:00, Балчуг
16:00 Создание "деревьев событий" от
День 3, 16:00-20:00, Балчуг
16:00 Анализ получившейся суммарной
День 4, 16:00-20:00, Балчуг
16:00 Идентификация на карте возможных сценарных

24.11 (день 1)
Присутствуют 24.11
Рабочая область
Basic
Maps
Present
Future
Stakeholders

25.11 (день 2)
Люди
День 2, 16:00-20:0
Рабочая область
Present and future wheel
Midcasting
Future - backcasting

26.11 (день 3)
Рабочая область
Результа
Сценарная карта
Stories
Фото

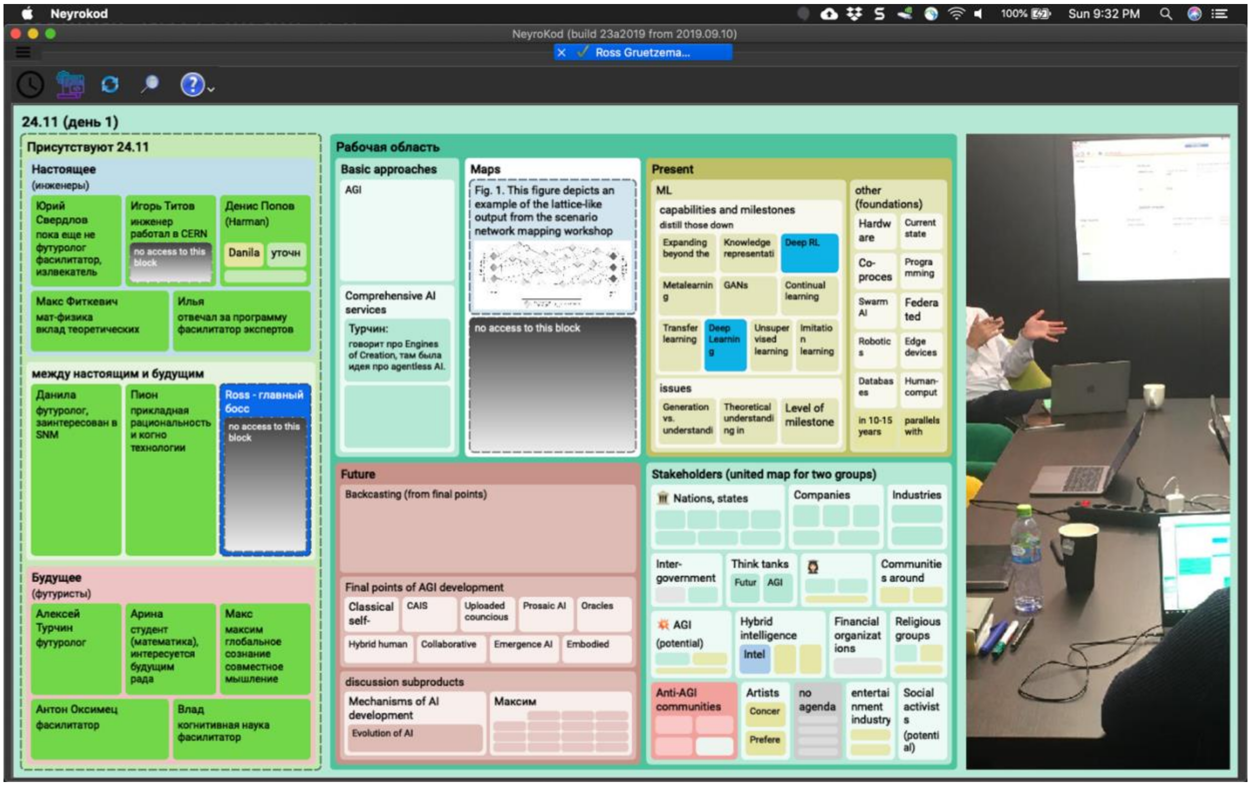
27.11 (день 4)
Единая карта AGI
Map AGI 1.nk
Один блок - для копирования
Добавления - задачи
pick a few plausible paths
prune the
New
move
Добавления - идеи
Уточнение проблемы -
no access to
Решение проблемы
no access to
Negative consequences
Фото
Люди
Пион
Данила

Постобработка
Ross
Теор
Surv
Yury Sverdlov, [27.11.19
Влад - хочется еще подумать в НейроКоде
Денис
можно
Максим
Арина
Надо
Илья
Данила

Figure 27: NeyroKod’s project management interface for the third SNM workshop.

Given the power of the software product being used during this workshop many elements were substantially different. Overall, the objective was to identify technical paths to AGI as well as to identify safety critical elements and to add a milestone layer. This workshop, unlike the two previous workshops, was held independent of any other events. Consequently, it was conducted over four days involving four half-day workshops as suggested by List (2005) and was thus the only one of the developmental workshops conducted for the broader study that involved a full workshop as List had instructed.

The workshop plan was made with the assistance of the NeyroKod development team. However, based on previous experience, it was clear that facilitation of the SNM workshopping process with this new technological capability would be something that would have to be fine-tuned onsite in Moscow during the workshop. Prior to the first half-day session a 30-minute meeting was held with the NeyroKod developers. The first day was a Sunday, and as a result, involved the highest number of participants of any of the days. This first day also allowed for the use of two rooms in the co-working space that was hosting the event. Each of the remaining days only allowed for a single conference room. So, the first day was the only day of this workshop series (or any of the developmental workshop sessions) to be conducted with two entirely independent groups. This was possible because several participants assisted with the facilitation of the process. However, prior to splitting up, the entire group first convened in one room where a 30-minute introduction was conducted. This was followed by going around the room and having the participants explain their background and their interests with regard to AGI. Following this introduction exercise the groups split into two and the second group went to the second conference room. Based on the experiences from the 1st and 2nd workshops, the groups were determined by the facilitator based on the participants' different areas of expertise. After splitting into the two groups, one group worked to identify present technologies or near-term future technologies, while the other group worked to identify desirable end states for backcasting. The lead facilitator went between the two rooms to monitor progress while co-facilitators assisted in the meantime. Following the identification of the present and future groups, the two groups reconvened, and, because of the limited time, the remainder of the workshop was used to identify a stakeholder map. The results from the first day of the workshop are shown in Figure 29.



24.11 (день 1)

Присутствуют 24.11

Настоящее
(инженеры)

Юрий Свердлов
пока еще не футуролог фасилитатор, извлекатель

Игорь Титов
инженер работал в CERN
no access to this block

Денис Попов
(Harman)
Danila
уточн

Макс Фиткевич
мат-физика вклад теоретических

Илья
отвечал за программу фасилитатор экспертов

между настоящим и будущим

Данила
футуролог, заинтересован в SNM

Пион
прикладная рациональность и когно технологии

Ross - главный босс
no access to this block

Будущее
(футуристы)

Алексей Турчин
футуролог

Арина
студент (математика), интересуется будущим рада

Макс
максим глобальное сознание совместное мышление

Антон Оксимец
фасилитатор

Влад
когнитивная наука фасилитатор

Рабочая область

Basic approaches
AGI
Comprehensive AI services
Турчин:
говорит про Engines of Creation, там была идея про agentless AI.

Maps
Fig. 1. This figure depicts an example of the lattice-like output from the scenario network mapping workshop
no access to this block

Present
ML
capabilities and milestones
distill those down
Expanding beyond the
Knowledge representati
Deep RL
Metalearnin g
GANs
Continual learning
Transfer learning
Deep Learnin g
Unsuper vised learning
Imitatio n learning
issues
Generation vs. understandi
Theoretical understandi ng in
Level of milestone

other (foundations)
Hardw are
Current state
Co-proces
Progra mming
Swarm AI
Federa ted
Robotic s
Edge devices
Databas es
Human-comput
in 10-15 years
parallels with

Future
Backcasting (from final points)
Final points of AGI development
Classical self-
CAIS
Uploaded councious
Prosaic AI
Oracles
Hybrid human
Collaborative
Emergence AI
Embodied
discussion subproducts
Mechanisms of AI development
Evolution of AI
Максим

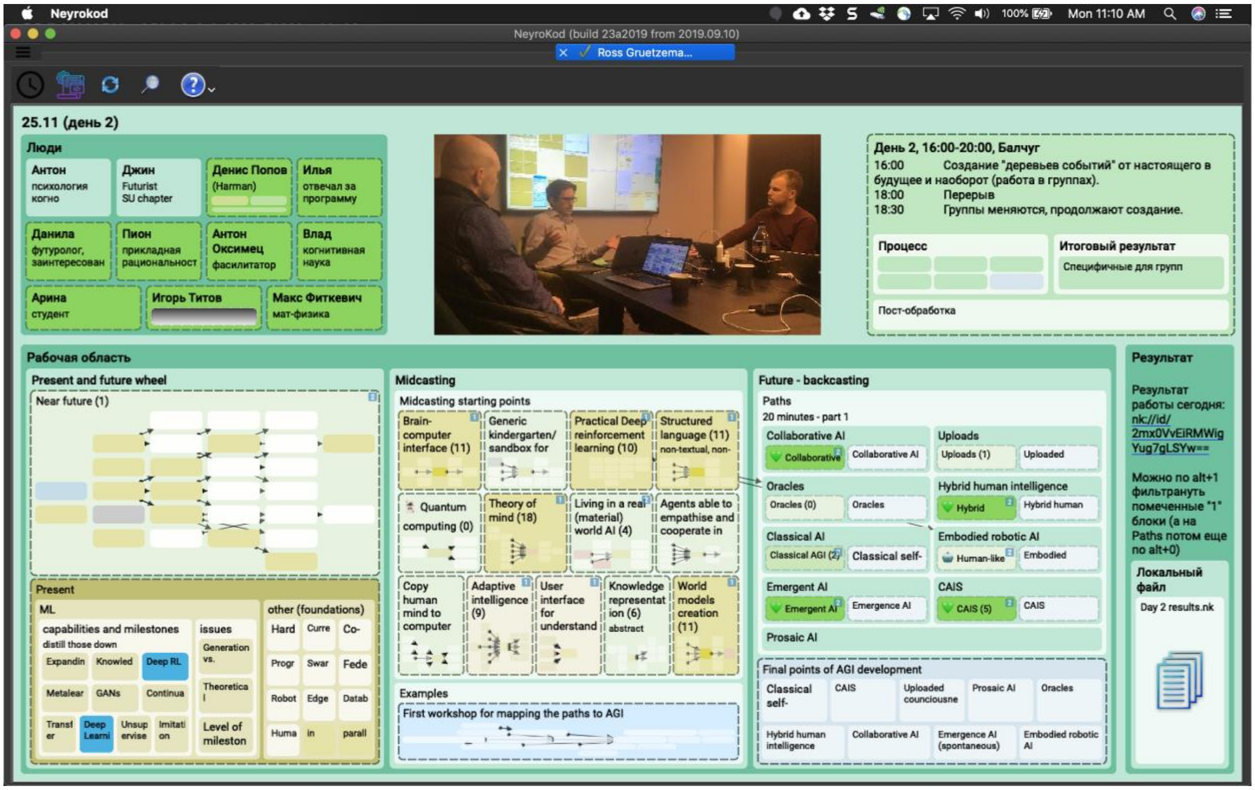
Stakeholders (united map for two groups)
🏛 Nations, states
Companies
Industries
Inter-government
Think tanks
Futur
AGI
Communitie s around
AGI (potential)
Hybrid intelligence
Intel
Financial organizat ions
Religious groups
Anti-AGI communities
Artists
Concer
Prefere
no agenda
entertai nment industry
Social activist s (potenti al)

Figure 29: The NeyroKod results from the first day of the workshop.

Progress during the second day of the workshop was slowed some because only one conference room was available so splitting into groups was not possible[116]. Initially, a futures wheel was used to generate topics for midcasting. This took half of the first two hours of the workshop. Following this, a session was held for midcasting from the results of the futures wheel, however, because no event trees had been generated in the previous day's process, a significant amount of time was required to explain this process and after the explanation and a demonstration an early break was taken. Following the break most of the remainder of the session was focused

[116] Separate rooms are necessary for splitting into groups when using NeyroKod unless the room being used is supplied with two projectors or monitors.

on midcasting. The session ended with an attempt to begin backcasting because that had also been scheduled for the second day. The results are shown in Figure 30.



25.11 (день 2)

Люди

Антон психология когно

Джин Futurist SU chapter

Денис Попов (Harman)

Илья отвечал за программу

Данила футуролог, заинтересован

Пион прикладная рациональност

Антон Оксимец фасилитатор

Влад когнитивная наука

Арина студент

Игорь Титов

Макс Фиткевич мат-физика

День 2, 16:00-20:00, Балчуг

16:00 Создание "деревьев событий" от настоящего в будущее и наоборот (работа в группах).
18:00 Перерыв
18:30 Группы меняются, продолжают создание.

Процесс

Итоговый результат

Специфичные для групп

Пост-обработка

Рабочая область

Present and future wheel

Near future (1)

Present

ML

capabilities and milestones

distill those down

Expandin | Knowled | Deep RL

Metalear | GANs | Continua

Transf er | Deep Learni | Unsup ervise | Imitati on

issues

Generation vs.

Theoretica l

Level of mileston

other (foundations)

Hard | Curre | Co-

Progr | Swar | Fede

Robot | Edge | Datab

Huma | in | parall

Midcasting

Midcasting starting points

Brain-computer interface (11)

Generic kindergarten/ sandbox for

Practical Deep reinforcement learning (10)

Structured language (11) non-textual, non-

Quantum computing (0)

Theory of mind (18)

Living in a real (material) world AI (4)

Agents able to empathise and cooperate in

Copy human mind to computer

Adaptive intelligence (9)

User interface for understand

Knowledge representat ion (6) abstract

World models creation (11)

Examples

First workshop for mapping the paths to AGI

Future - backcasting

Paths

20 minutes - part 1

Collaborative AI: Collaborative | Collaborative AI

Uploads: Uploads (1) | Uploaded

Oracles: Oracles (0) | Oracles

Hybrid human intelligence: Hybrid | Hybrid human

Classical AI: Classical AGI (2) | Classical self-

Embodied robotic AI: Human-like | Embodied

Emergent AI: Emergent AI | Emergence AI

CAIS: CAIS (5) | CAIS

Prosaic AI

Final points of AGI development

Classical self- | CAIS | Uploaded counciousne | Prosaic AI | Oracles

Hybrid human intelligence | Collaborative AI | Emergence AI (spontaneous) | Embodied robotic AI

Результат

Результат работы сегодня: nk://id/2mx0VvEiRMWigYug7gLSYw==

Можно по alt+1 фильтрануть помеченные "1" блоки (а на Paths потом еще по alt+0)

Локальный файл

Day 2 results.nk

Figure 30: The NeyroKod results from the second day of the workshop.

The third day of the workshop began with less progress having been made than anticipated because backcasting had only begun at the end of the previous day's session. The lack of progress from day two's efforts on backcasting is evident in Figure 30. Thus, the first half of the day was spent on continuing backcasting. This was a challenging task given the lack of technical expertise and the large number of participants. The group was split into three smaller groups for the event tree creation process, however, some groups were more effective than others due to the different levels of expertise. This made facilitation challenging for the two groups lacking substantial technical expertise because each was in need of ongoing guidance. This process was continued until the half-way break, and then for an hour after that. For the final portion the group reconvened and discussed the next steps involving the connecting of the event trees. No progress was made during this time, and the decision was made to end early and let NeyroKod experts connect the trees in the time before the final workshop the next day. The results for day three are shown in Figure 31.

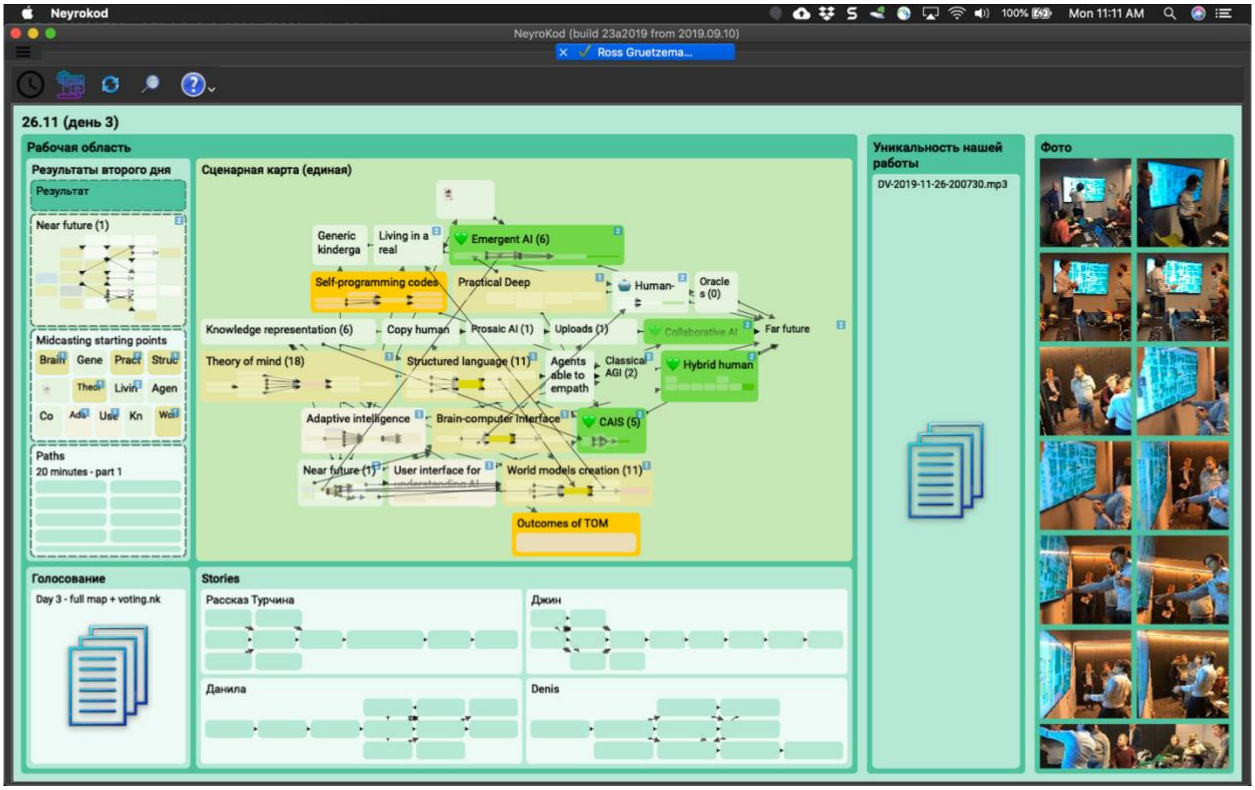


Figure 31: The NeyroKod results for the third day of the workshop.

At the start of the final day of the workshop all of the event trees that had been previously created had been connected. However, many of the connections were less than ideal. So, the fourth workshop was dedicated entirely to cleaning up the existing map and to identifying missed items so that they could be added to the final version of the resulting scenario network map. This process was conducted with everyone convened as a single group and was poorly organized because it involved significant discussion among the most prolific contributors to the map. The resulting map was considered a major success; however, much room for improvement remained as well. The final scenario network map is shown in Figure 32.

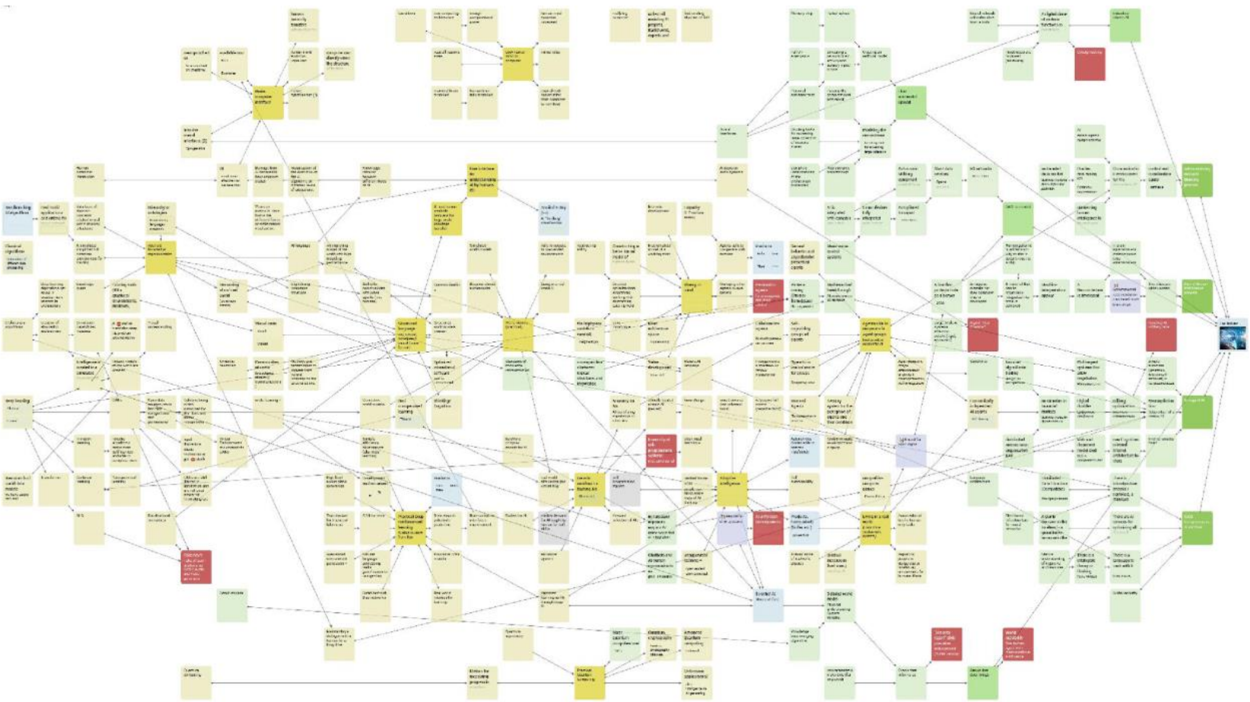


Figure 32: The final scenario network map from the fourth day of the third SNM workshop. This map has expanded the event trees into a single map. AGI is shown on the far right with a blue icon on the node. The red nodes represent risks from AI technologies.

Participants of this workshop came from a wide variety of backgrounds and were starkly different from those from any of the other workshops that were conducted (see Table 12). There was a scarcity of technical experts, and only one participant with substantial technical expertise participated in at least three out of four of the days. The two experts with the most technical expertise were only able to participate in one day of the workshop each, and these were different days so their expertise was not very influential on the outcome of the overall resulting map. All participants had a thorough knowledge of the different aspects of AGI and risks associated with such a system that has values poorly aligned with those of humans, but there was not consistent knowledge of the work of the AI strategy community as most of the participants were members more broadly concerned about positive and negative futures for humanity. NeyroKod is intended for mapping futures to explore such possibilities, and the participants were mostly all very familiar with this software. Thus, in some sense the participants' common shared expertise could be thought to be the mapping of possible futures.

The common expertise of the participants, and the lack of substantial technical expertise, made this workshop interesting for exploring the possible uses of the technique for strategic planning purposes, as is suggested by the holistic forecasting framework proposed in this study. Because the common expertise was not centered on topics related to strategic planning it likely made exploring this element more challenging, but, despite this, the experience seemed to support the viability of the technique for this purpose, particularly when the NeyroKod software is being used. The challenges of the final day, in connecting the event trees and identifying missing items for further decomposing into more event trees, would have been easier with a more consistent expertise among participants. However, the fact that a map was generated in its entirety with different participants in different sessions over multiple days suggests that there is much value for using the technique in tandem with NeyroKod for strategic planning applications. Furthermore, such strategic planning applications likely are not limited to technology forecasting efforts, and the technique in combination with the software would likely be beneficial for a wide variety of different applications in business.

A survey was also administered to participants of this workshop as well. The response rate was high, with 10 of 12 participants completing the survey. Two participants attended for one day, two participants attended for two days, two participants attended for three days and six participants attended for each of the four days. Participants of this workshop series, like participants of the first SNM workshop series, were overwhelmingly pleased with the experience as well as the resulting map. While most did not have expertise related to AI and had been expecting a larger group of AI experts to be in attendance than was the case, they perceived the method to be well suited for both AI purposes as well as other purposes. Furthermore, they felt that the map that results from this workshop, despite the lack of AI-specific expertise, was useful and could be further refined in ongoing efforts.

These participants also had some constructive comments in the survey. A common comment was the need for participants to be familiar with the technique and software in order to optimize the results. This is a major factor, because, especially when using NeyroKod, participants must be able to effectively collaborate in the digital environment as well as in the room with the other participants directly. It is actually challenging to balance this, and this suggests that more work should be done to determine specific guidelines for conducting the SNM workshopping processing using NeyroKod. Perhaps because surveys were completed on the last day, and the events of the last day were thus most fresh in the minds of the participants, there were numerous comments on the challenges of unifying the map after creation of the different event trees. It was during this portion of the workshop, on the last day, where there were the most difficulties in reaching consensus (this seemed more challenging due to one large group and due to the combination of work being done digitally with direct conversations between participants; the latter made it difficult to keep up with the numerous different elements being worked on). For this reason, and for other examples described in the survey results, participants seemed to agree with the previous conclusion that more work needs to be done developing a specific set of guidelines for using the method with NeyroKod. Despite this, the facilitator and participants felt that there was increased value conducting SNM with it.

6.6 Judgmental Distillation Mapping

JDM requires an input of numerous other forecasting techniques for experts to iterate over during the interviews and questionnaires. However, the results from the methods discussed in previous sections of this chapter were not all used for this purpose because their analysis was not entirely complete at the time when interviews began. Initially, only the mixed methods study, the AI practitioner survey and the simple bibliometric analysis were complete when the first round of JDM interviews began. The results from the survey and some qualitative results from the mixed methods interviews were used to inform the first round of interviews. The results from the SNM workshops that had been completed at the time[117], as well as the extrapolation of the DeepMind financial data, was used to inform the second and third rounds of interviews.

6.6.1 JDM Iterative Interviews

The JDM process involves iterative rounds of interviews or questionnaires with AI experts. For the demonstration of this technique in this study only interviews were used. The first and second rounds of interviews consisted of formal interviews, while the third round of interviews was more informal in nature and conducted during poster presentations at AI conferences[118]. The interviews were conducted at a total of 16 of the most significant AI conferences from December 2018 to December 2019.

[117] Because workshops were being conducted in between different conferences, the results from the first workshop were used during the second round of interviews, and the results from the first and second workshops were used for informing the 3rd round of informal interviews.

[118] After attending a large number of academic conferences for the mixed methods, first round JDM and second round JDM interviews, it was determined that approaching participants to speak about the more granular details during the periods when they were already available to discuss their work would be a suitable method for obtaining more information from more experts for the third round of interviews. One reason that this was even better for this final round of interviews is a result of the fact that this the last round of JDM interviews was primarily concerned with more granular details of current or emerging AI technologies (in order to identify forecastable targets for the nodes in the JDM which had been generated during the first two rounds of interviews).

6.6.1.1 Expert Selection

Experts were selected using different AI conferences, some more general AI conferences and other more specialized AI conferences, in order to curate a sample that best represented the population of AI researchers. The first round of interviews was conducted at two conferences, NeurIPS[119] and AAAI[120], during the winter of 2018/2019. The second round of interviews was conducted during the early summer of 2019 at six conferences: ICLR[121], AAMAS[122], ICRA[123], NAACL[124], ICML[125] and CVPR[126]. The third and final round of JDM interviews was conducted at 7 conferences between late summer and early winter of 2019: AGI-conf[127], IJCAI[128], ICCV[129], INLG[130], CoRL[131], EMNLP[132], IROS[133] and NeurIPS.

One major concern with the selection of conferences to be used for the rounds of interviews was sample bias. The first round only included two conferences, which was less than ideal, but there were not enough major AI conferences scheduled during winter 2018/2019 to allow for attending a wide enough variety to have adequately addressed this concern. However, each of the conferences were very large conferences that included all different types of AI research. The first of these was NuerIPS, held from December 3rd-8th, 2018. The other conference in this first round

[119] The annual conference on Neural Information Processing Systems in Montreal, QC, Canada.

[120] The annual conference for the Association for the Advancement of Artificial Intelligence in Honolulu, HI, USA.

[121] The International Conference on Learning Representations in New Orleans, LA, USA.

[122] The annual conference on Autonomous Agents and Multi-Agent Systems in Montreal, QC, Canada.

[123] The International Conference on Robotics and Automation in Montreal, QC, Canada.

[124] The annual meeting of the North American chapter of the Association of Computational Linguistics in Minneapolis, MN, USA.

[125] The International Conference on Machine Learning in Long Beach, CA, USA.

[126] The IEEE conference on Computer Vision and Pattern Recognition in Long Beach, CA.

[127] The annual conference on Artificial General Intelligence in Shenzhen, China.

[128] The International Joint Conference on Artificial Intelligence in Macau, China.

[129] The International Conference on Computer Vision in Seoul, Korea.

[130] The International Conference on Natural Language Generation in Tokyo, Japan.

[131] The Conference on Robot Learning in Osaka, Japan.

[132] The conference on Empirical Methods in Natural Language Processing in Hong Kong, China.

[133] The International Conference on Intelligent Robots and Systems in Macau, China.

of interviews was AAAI, held from January 27th to February 1st, 2019. These are each two of the top five major AI conferences that involve a wide variety of different AI topics, from theory development to computer vision to NLP.

The large number of conferences in the second and third rounds was necessary in order to mitigate sample bias. Each round included conferences specific to the AI subdisciplines of robotics, computer vision and NLP. Each also included one of the two leading AI conferences (both of which focus on deep learning).

6.6.2 Judgmental Distillation Maps

In this section we first review the computation of the probabilistic forecasts in the internal nodes and demonstrate this using fictitious distributions in lieu of expert elicited forecasts. We then review the results of the JDM process used to generate a judgmental distillation map using expert elicitation.

6.6.2.1 Monte Carlo Computation

Section 5.1 discussed the foundational concepts of the JDM technique but did not demonstrate how the Monte Carlo simulation would be practically applied. Because the JDM process was only demonstrated qualitatively[134], and the group elicitation techniques proposed for determining the input probabilistic forecasts for the Monte Carlo simulation were not conducted, we are unable to demonstrate the technique in a practical application. However, the technique is useful and the only known scenario mapping technique to generate probabilistic forecasts for future scenarios.

Figure 8 in Section 5.1 depicts a simplified example of a judgmental distillation map. A slightly expanded map is shown in Figure 33. This represents the state of the map after the first round of JDM. Two levels of general intelligence precede human level AI because many experts

[134] Due to resource constraints, discussed previously.

anticipate arriving at very powerful systems that remain very far from HLAI or notions of AGI, and these systems vary significantly[135]. One significant difference between this figure and Figure 8 is the inclusion of hypothetical probabilistic forecasts for the external nodes. The normal distribution was used for these demonstrative distributions because it is more intuitively perceived than alternatives, like the Gamma distribution, which may be more appropriate for technology forecasting due to its being right skewed with a fat tail. It is also more easily parameterized and parameterized in a manner with which more readers are likely familiar (i.e., with the mean and variance/standard deviation).

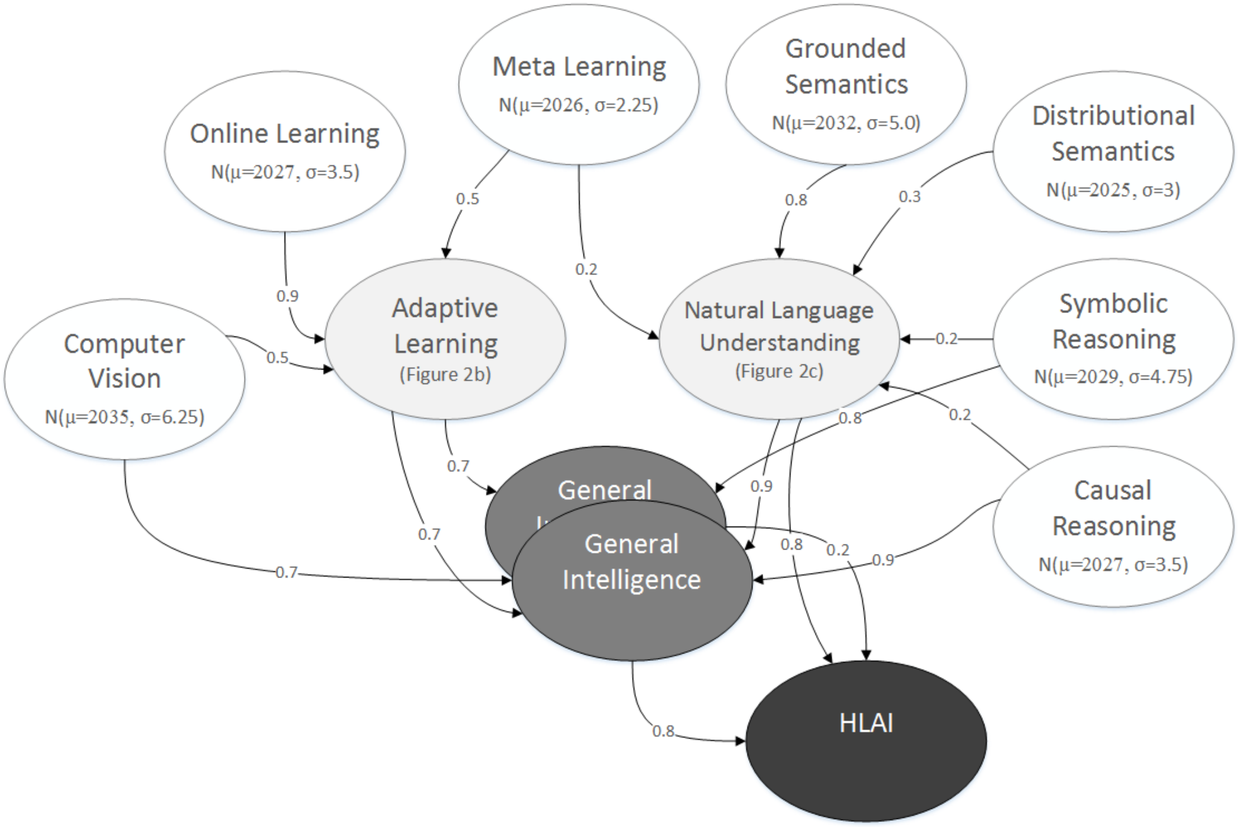


Figure 33: This is an illustrated example of the judgmental distillation map that was obtained after the first round of JDM. Hypothetical distributional forecasts are included for the external nodes.

[135] A Twitter exchange between Yann LeCun and Eliezer Yudkowsy in December 2019 illustrates this possibility – https://twitter.com/ylecun/status/1204038764122632193.

Using the hypothetical distributional forecasts for the external nodes we can compute the distributional forecast for the internal nodes. As this is intended solely for demonstration, the demonstration is limited to the first level of internal nodes which are denoted in Figure 32 by a light grey shading. Figure 34 depicts the probabilistic forecasts for each of these nodes. Each of these generated distributions was generated with n = 10,000 and can be seen to not be smooth like parameterized distributions.

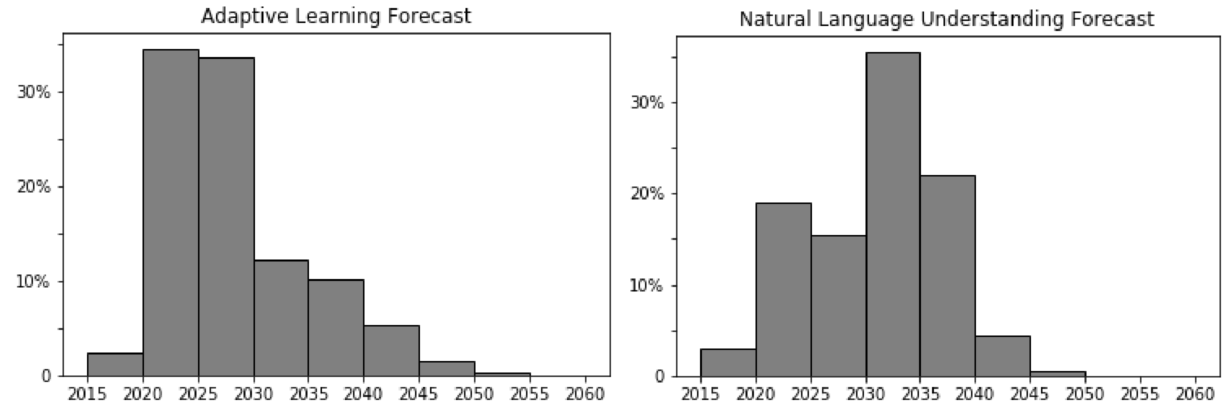


Figure 34: a) The probabilistic forecast for the adaptive learning node computed using Monte Carlo simulation and b) the probabilistic forecast for the natural language understanding node computed using Monte Carlo simulation

6.6.2.2 Final Judgmental Distillation Map

The final map did not vary that much from the initial map generated in the first round. Some additions and modifications were made during the second JDM round, but during the third round the informal discussions were only used to identify the weights for the different causal paths leaving the near-term plausible nodes. Each of the rounds involved different approaches to eliciting the opinions of experts regarding the information of interest. These are described below.

The first round of interviews focused on identifying the next generation technologies that would comprise the internal nodes and lead to RTAI (e.g., HLAI or CAIS). These interviews focused on experts' areas of expertise but also inquired regarding their opinions on some of the topics which, based on the earliest interviews, seemed to be recurring most frequently with respect to their potential to transform society. These recurring topics were included in the working map that was developed as the interviews were ongoing. Not all of these interviews explicitly mentioned RTAI (e.g., notions of AGI) because some interviewees' answers to early questions suggested more value could be extracted by avoiding these topics as they were unlikely to respond constructively[136]. Aggregation of the data collected from these interviews into a judgmental distillation map requires judgement on the part of the forecaster[137], particularly when generating the initial map. Figure 8 and Figure 34 depict the initial map from this first round at an early state and a final state, respectively.

The second round of interviews again involved asking primarily questions about experts' areas of expertise. Unrelated topics were not typically included, however, the experts approached and interviewed during this round generally worked on research that had implications for numerous of the elements in the initial map. After initial questions, experts were shown the map and given time to process and understand this. Experts were then asked for input regarding whether they thought something was missing or out of place. Most experts responded that the map looked roughly correct, although some details were out of place. Because the map generated during the initial round was very generic and lacked specificity this may have enabled experts to agree whereas a finer-grained map would have precluded agreement. However, agreement regarding the content of the map is desirable because, in order to generate quantitative forecasts, a number of

[136] As noted earlier with respect to the survey, many AI experts have very strong opinions on the grand goal of AI research and different notions of AGI.

[137] This was mentioned in Chapter 5, and it is crucial that the forecaster have both technical AI expertise as well as forecasting expertise. Future efforts to generate practical forecasts should consider teams of persons with expertise in each of these areas.

plausible paths is necessary, and too many would likely require many more external nodes for which more forecasts would have to be collected if JDM was operationalized. Thus, we found the first two rounds successful. One observation that could be useful for future implementations of JDM is having experts who participated in the early interviews of the second round review the final scenario map, because the final map included significantly more external nodes that the early map, and it would be interesting to know if these were all necessary or if some pruning could be conducted.

The third round of informal discussions was significantly different from the previous two rounds, foremost because there were no formal interviews. This increased interview quantity at the cost of quality. In general, all of these interactions were short and commonly required segueing from the experts' posters' content to a next generation technology (related to their active research foci). This commonly only allowed for eliciting information regarding the weight of a single causal path in the scenario map from each interaction, and for some causal paths only a single elicitation was possible. Moreover, it was commonly difficult to frame questions such as to directly elicit the value of the causal paths due to the circumstances and not displaying the map during the interaction. For example, some questions were formulated like "how significant a role do you anticipate meta reinforcement learning playing in the development of generalizable model-based reinforcement learning?" Such questions would elicit qualitative answers, and follow-up questions would try to elicit more detailed information to varying degrees of success. This round was not as successful as we had hoped[138]. This led to not being able to conduct any decomposition, yet it was

[138] This new approach was employed because it became more difficult to schedule interviews with relevant experts as the study progressed further. For one, many of the most relevant experts who were willing to be interviewed had already been interviewed or solicited for an interview by the time of the 3rd round of the JDM process. This is a consequence of there being a very limited number of experts in the world and the very selective process followed when identifying experts to invite for participation. This limitation didn't become apparent until this final round of JDM.

a learning opportunity, too, and the experience suggests that quality is prefered to quantity in JDM when possible. Delphi-like questionnaires were not used; this may be an excellent option for any of the three rounds and should be considered in the future.

The final scenario map generated by the JDM process in this study is depicted in Figure 34. It only includes two more external nodes and two more first internal layer nodes that the scenario map resulting from the first round of interviews (the map did not change between the second and third rounds, only the weights were added). Some nodes were rearranged during the second round, which is apparent when comparing Figure 34 with Figure 35. Moreover, some nodes were heavily modified during the second round with feedback from experts about how they may actually contribute to the next generation technology in combination with some of the other incoming technologies. Overall the demonstration was deemed successful, and a more detailed discussion is included in the following chapter.

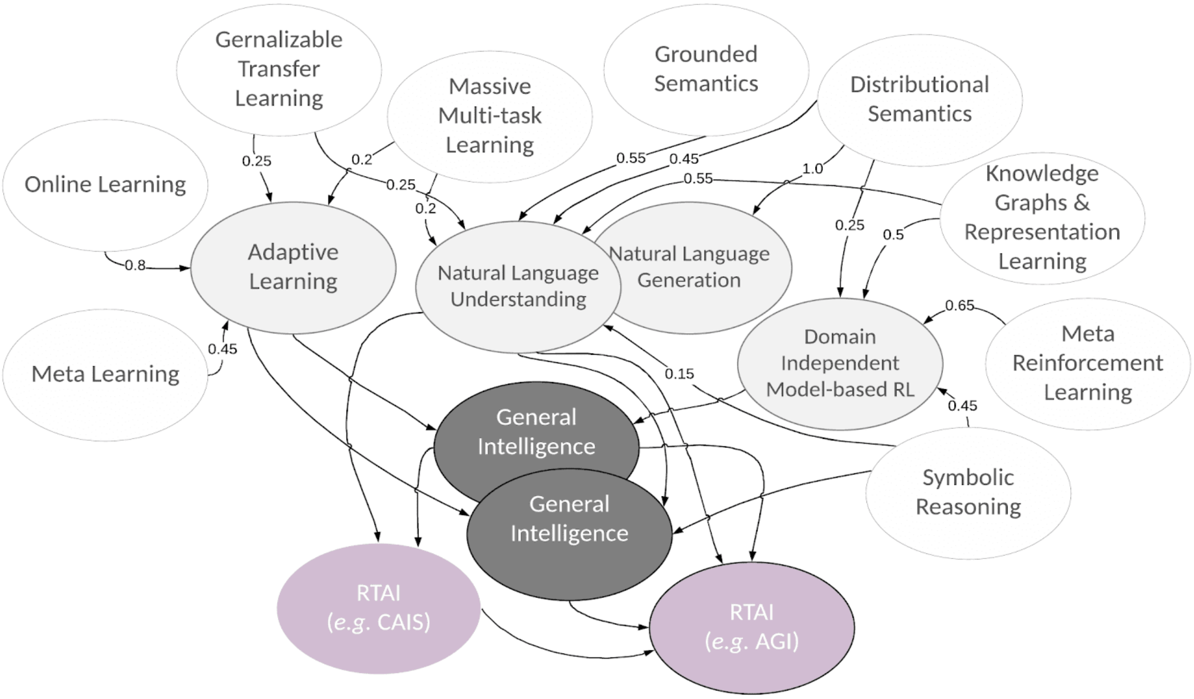


Figure 35: The final judgmental distillation map after three rounds of interviews and discussions with experts. The lavender nodes depict RTAI and end states. The grey shaded nodes likely represent DTAI, but this is not certain.

7 USING THE DELPHI: A RESEARCH AGENDA

Many components of this study represent the first attempt to use these techniques for the purpose of forecasting AI progress[139]. However, the objective of this study is not to generate actionable forecasts, but rather, to critically explore the different techniques that can be used to do so. Thus, rather than using the Delphi technique for generating a forecast, the Delphi technique was used to generate a research agenda[140] for AI forecasting. This was done in order to evaluate an adaptation of the policy Delphi (Turoff 1970), and previous studies that have used the Delphi for generating questions and topics of interest in a specific research domain, for generating questions and topics of interest in the context of AI forecasting. For the purpose of completing the study as originally proposed, this variation was chosen instead of applying the Delphi for quantitative forecasting because it allowed for exploring the possibility of using the Delphi for generating questions and forecasting targets using experts. This application of the Delphi could be useful for adapting Tetlock's (2017) full-inference-cycle tournament framework for use on TAI and DTAI forecasts.

[139] The research agenda – and this Delphi study – was directed at forecasting AI progress rather than forecasting TAI. There are numerous reasons for this. For one, forecasting AI progress is more acceptable for academics not involved in AI strategy or AI safety research. Also, there is significant value to any work on forecasting AI progress, much of which may also be valuable for forecasting TAI. Moreover, there is utility in near-term forecasts of AI technologies which is largely ignored by this study. Furthermore, limiting the Delphi process responses to a poorly defined concept would be an ineffective use of experts' time as it would likely lead to limited or biased results.

[140] The primary focus is evaluation of the technique for future applications in forecasting such as the generation of forecasting targets for prediction markets or forecasting tournaments. Consequently, the results of the study and the research agenda itself are not discussed at great length in this text. Interested readers can see Gruetzemacher et al. (forthcoming). Also, the results and the research agenda outline are included in Appendix F and Appendix G, respectively.

The methods used in this Delphi process were outlined in section 4.2. The questions to be asked were also included in that section in Table 3. Invitations were sent to 32 different experts with expertise on a broad range of topics related to forecasting AI progress. 15 of these experts responded, including participants from a leading AGI lab, from the world's leading universities[141], from both the European and US governments, as well as others who have conducted work on the topic which has been highly cited. This chapter reports the details of the process that was described in Chapter 4, as well as the details of the analysis of the data that was collected from the experts. The results are then reported in the next section of the chapter. They are presented in an objective manner, without any modification from their original form. The final subsection of the chapter is dedicated to identifying things that, based on the other observations and techniques examined in this study, may have been overlooked or discounted by the experts based on the raw results reported from the Delphi data.

7.1 Notes from the Delphi Process

The Delphi process conducted here was an adaptation of the policy Delphi (Turoff 1970) and other previous work (Dahmen et al. 2013, Kellum et al. 2008, Gordon and Barry 2006, Dimmit et al. 2005) specifically designed for the purpose of identifying the salient research questions and methods for forecasting AI progress. The invited experts' areas of expertise included economics, technological forecasting, AI forecasting and AI foresight, and they were representative of academia, government, industry and nonprofit organizations. A breakdown of the responses and non-responses by employer type is shown in Table 13. In the table it can be seen that the responses are representative of all of the different employer types.

[141] e.g., faculty, research scholars and PhD students from the University of Oxford, the University of Cambridge and Johns Hopkins University.

Table 13: Responses and Non-Responses by Employer Type

	Response	No Response
Academia	5	7
Government	2	0
Industry	4	4
Nonprofit	4	6
Total	15	17

Of the fifteen respondents who participated, ten had published on topics related to forecasting AI progress. These respondents had a mean and median of 5,977.6 and 1,195 citations, respectively, with a mean and median h-index of 22.3 and 11, respectively. All of the non-respondents had published on the topic. They had a mean and median of 7,455.3 and 2,088 citations, respectively, with a mean and median h-index of 21.1 and 12 respectively. Based on this data, it appears that the sample of experts are not significantly less qualified or accomplished than the non-respondents. Consequently, the respondents are deemed to be a representative sample of experts on forecasting AI progress, and the research agenda is thought to be based on objective expert opinion.

The first round of the Delphi involved the questionnaire, depicted in Table III. The first question was designed to get a sense of the tractability of the problem from the experts. The second question elicited opinions about the most important questions and the third question elicited opinions about the most important methods to be used. The final question asked about neglected topics in order to try to identify topics that may have been overlooked by participants when answering the second question.

Following the collection of responses from the questionnaire, a week was spent aggregating the results into a brief summary, a list of the top questions and a list of the top methods. Each response was included in the lists of questions and methods, and no items were excluded. However, substantial effort was made to combine similar topics, even when there were major differences in how the topics were expressed. For some items that were outliers, for which no other items were considered similar, they were rephrased in a manner to be consistent with what was perceived to be meant by them based on the facilitator's knowledge of each participant's specific research interests. The aggregated results from the Delphi questionnaire following the first round are included in Appendix F.

These results were reported back to the participants and participants were given the opportunity to contribute comments and discuss the responses from the questionnaire. Because the lone topics were frequently unique, and sometimes esoteric, it was difficult to rephrase some of them so that they would make sense in a few words to a broad group of expert forecasters. One participant left two comments, each regarding clarification of different elements in the list. Disappointingly, no discussion of the topics was conducted and no other participants helped to clarify the topics that received the comments.

For the second round, the lists were then added to a spreadsheet so that participants could score the different questions and the different methods. A spreadsheet was used in lieu of a survey or online form because of the number of different questions being asked. Online forms or surveys would have required a large number of different pages, which may have led to a lower response rate. The spreadsheet was used so as to maximize the response rate by compacting the required work of scoring 67 items to a single page. The spreadsheet also enabled the elicitation of two scores for each question: one for importance and another for feasibility. This would have required two separate questions with the different online form and survey options that were explored.

12 of the 15 respondents completed the scoring element of the 2nd round of the Delphi, which is roughly in line with the best second round response rates from previous studies (Gordon and Barry 2006). However, the majority of participants did not complete all items for each of the different dimensions being scored. This was expected, but the severity of the problem was much greater than had been anticipated.

7.2 Imputation

Because of the large amount of missing data due to the numerous incomplete values in the spreadsheets that had been distributed to the experts, the raw results could not be used directly to aggregate the scores for the questions and the methods. Instead, multiple imputation (Rubin 2004) was required in order to approximate the missing data so that the results could be reported.

Given the existing values, multiple imputation predicts the missing values using Bayesian ridge regression and by randomly sampling from the learnt posterior distribution. This is conducted for a large number of samples, which, in this case, was 10,000. These results are reported in Appendix F. Respondents were instructed to score each question and method on a scale of 1 to 5, however, not all respondents used a discrete scale. Consequently, the medians that are reported in Appendix F are not all whole numbers.

7.3 Results

The full results are too extensive to report in the body of this document and are thus included in Appendix F[142]. The top scoring questions are shown in Table 14. Considering that these top five questions are for forecasting AI progress, and not transformative AI specifically, it seems that much of the work conducted in this study is still directly relevant to some of the topics that experts

[142] Gruetzemacher et al. (forthcoming) report the results of this in more detail. Interested readers are encouraged to review this as well.

feel to be most important. Particularly interesting is the inclusion of how to use forecasts to inform decision makers and the need to identify plausible paths to TAI technologies.

Table 14: Top Question Scores

Question	Imputed Mean	Median
What questions/targets matter for practical, near-term decision making?	4.21	5
How do we utilize forecasts to inform decision makers and develop appropriate and measured initiatives/interventions?	4.14	5
How do we best validate forecasts of AI progress: historical data/near-term progress?	4.12	5
How do we identify the most plausible paths for a variety of transformative AI technologies/systems?	4.03	4
How can we decompose abstract AI technologies into more easily forecastable targets?	4.02	4

The top scoring methods are shown in Table 15. There was less disagreement about the different methods, as can be seen by the lower imputed mean scores for the highest-ranking methods. Hybrid methods and extrapolation, two of the focus topics of this study are in the top five methods found to be most important by the experts. This study also stresses the importance of identifying clear and effective forecasting targets. The first and last items in the top five methods have not been mentioned in this study. However, some of the questions being explored in interviews in the JDM process may constitute in-depth analysis of specific questions. This class is very broad, and it is not necessarily clear what specific questions are implied. Blue-team/red-team refers to a technique widely used in military or cyberdefense applications, for evaluating the robustness of a system or defense. In the context of forecasting this can be thought of as applying to the evaluations of the robustness of a specific forecasting model via this approach.

Table 15: Top Scoring Methods

Method	Imputed Mean	Median
In-depth analysis of specific questions	4.03	5
Hybrid methods (combining judgmental and statistical)	3.92	5
Identifying clear and effective forecasting targets	3.85	4
Extrapolation	3.77	4
Blue-team/red-team	3.75	4

The remainder of this section reports the results for the questions of interest and the most important methods in more detail. The questions/topics are reported first, then the methods are reported.

7.3.1 Questions/Topics of Interest

As noted previously, the full results can be found in Appendix F. Here, these results are briefly highlighted. This is done by grouping the different questions into clusters of questions and sorting these clusters by importance based on the scoring. These clusters are able to be grouped into three distinct groups: high-level topics, methods-related topics and dissemination of forecasts (just a single cluster that did not fit in any of the other groups). A general outline is shown below:

1. High-level topics
    a. Identifying the most important forecasting questions/targets
    b. The implications of timelines and evaluating AI progress
2. Methods-related topics
    a. Identifying useful indicators
    b. Modelling progress in AI
    c. Scenarios for AI development
    d. Improving forecasting efforts for AI

    e. Effectiveness of long-term forecasting
3. Dissemination of forecasts
    a. Identifying the best means for reporting forecasts to decision makers

7.3.2 Priority Methods

Not enough responses were elicited to compare the scoring of judgmental methods versus statistical methods because respondents appear to have perceived these topics as section headers and not items for scoring. This was not the intention, however, general observations from the results of the first round of the Delphi are able to be reported regarding the prioritization of these two classes of methods. It was found that three primary perspectives emerged regarding the methods used to forecast AI progress:

- Statistical modeling using indicators or metrics for measuring AI progress (~60%)
- Judgmental forecasting techniques for exploring plausible paths forward and for eliciting probabilistic forecasts (~25%)
- Hybrid methods, which use elements of both statistical and judgmental forecasting techniques (~15%)

Given the results reported above, this ordering was retained for reporting the different groups of statistical, judgmental and hybrid methods. Clusters were not used for the methods as they were for the questions, but the most important techniques were identified. These techniques are reported in the outline below:

1. Statistical forecasting techniques
    a. Measures and indicators
    b. Tech mining and data science
    c. Modelling
        i. Extrapolative modeling
        ii. Bayesian modelling
        iii. Machine learning

      iv. Simulations

2. Judgmental forecasting techniques
   a. Forecasting targets
      i. Identifying targets
   b. Methods
      i. Blue team/red team
      ii. In-depth analysis of specific questions
      iii. Simulation and role-play games
      iv. Scenario analysis
      v. Immersive observation
3. Hybrid forecasting techniques
   a. Expert adjustment
   b. Bayesian networks

7.4 Overlooked Items for a Research Agenda

The results and observations from this study suggest that there are still gaps in the resulting research agenda for forecasting AI progress. These include the development of training modules and calibration techniques for use with different methods of expert elicitation as well as simple things, for example: are forecasting techniques useful for applications other than technology forecasting as effective for technology forecasting?; are technology forecasting techniques as effective for forecasting different types of AI as they are for technology forecasting?; which existing judgmental forecasting methods are best suited for technological forecasting applications?

Based on observations in this study, one thing that was ranked the lowest item by the experts is something that seems to be very promising for any expert elicitation techniques: calibration training. This helps to reduce overconfidence and other cognitive biases, and it generates data that can be used for improving the aggregation of the results. It is central to techniques such as those described for superforecasting (Tetlock and Gardner 2016), and it is surprising that experts did not rank it higher.

The other overlooked topics concern questions of whether standard forecasting techniques apply to technological forecasting, and, more specifically, to AI forecasting. Strangely enough, the literature review did not identify sufficient existing efforts to answer questions concerning these issues. In general, while the Delphi technique is widely used in organizations, very little work has been done to evaluate its effectiveness for expert elicitation to generate forecasts (Rowe and Wright 2001). This is likely due to the fact that experts' time is hard to come by, and when the Delphi is used with experts the resulting forecasts are more valuable to organizations than academics. Thus, much work to assess the validity of the Delphi has involved students. Green et al. (2015) have found prediction markets to be more effective than the Delphi in the majority of situations, however, the Delphi is still best suited for many applications. Tetlock and Gardner (2016) further find that forecasting tournaments with superforecasting techniques applied for team selection and aggregating the results perform better than prediction markets for forecasting geopolitical events[143].

The examples in the previous paragraph demonstrate the lack of evidence for the effectiveness of forecasting techniques in applications relevant to technological forecasting. This should be a primary concern for those working on forecasting AI progress, because much effort could be wasted pursuing methods that are ineffective or suboptimal in the context of AI forecasting, but which work well for all other topics. Moreover, this concern should not be limited to just judgmental forecasting techniques but should also consider statistical and hybrid forecasting

[143] However, personal communication with Tetlock suggests that he feels these results would not hold for technology forecasting, and particularly AI forecasting. For this task he supports decomposition of complex forecasting targets and methods that incorporate expert opinion in forecasting tournaments.

techniques. Thus, it is recommended that identifying the most effective methods for forecasting AI progress be a priority topic for future work (as was suggested by experts, though not a prioritized topic); this can be most easily accomplished by empirical evaluations of existing techniques in the context of forecasting AI progress. Moreover, numerous comprehensive literature reviews may uncover significant work that sheds light on some of these issues but that is not widely known and is not reported in a major academic publication.

8 DISCUSSION

This study involved the demonstration of numerous forecasting techniques including a broad, multi-faceted framework forecasting TAI/DTAI/RTAI. The results were presented in the previous section, however, there is a lot to unpack regarding the implications of the data that was presented. In this section we address most of the major components of the study: the survey, the extrapolation, the mixed methods analysis, scenario network mapping (SNM), judgmental distillation mapping (JDM) and the holistic forecasting framework. The Delphi results we address at length in Chapter 9.

8.1 AI Practitioner Survey

The AI practitioner survey was the first major component of this study. Surveys have been widely used for forecasting different notions of AGI (Michie 1972, Baum et al. 2012, Grace et al. 2018) as well as for different applications of technological forecasting (Nemet et al. 2017, Baker et al. 2009; etc.). For this reason, it was critical to consider this technique as a baseline for forecasting TAI/DTAI/RTAI.

8.1.1 Implications of Results

The results of the survey have significant implications for future of work researchers as well as for other forecasts concerning the development of RTAI. For future of work researchers, the results suggest that extreme labor displacing AI scenarios should not be assumed to be far in the future (Brynjolffson et al. 2018) and thus ignored. For those who continue to use surveys for forecasting RTAI, the methods utilized in this study can be a good guide to consider in the design of future survey instruments.

The most significant implications are those for future of work researchers because existing work on the topic does not consider different levels of TAI that could be precipitated by discontinuous technological change. While this portion of the study does little to explore the mechanisms through which such discontinuous progress could occur, it does illuminate the possibility of impacts to labor markets more severe than what previous studies have suggested. For example, it suggests that those working on technologies that could produce discontinuous progress anticipate a median 50% of chance of this occurring in 15 years[144]. Such levels of labor displacement cannot possibly be modeled with the techniques currently being promoted for such models which rely on historical data (Das et al. 2020). While it has been suggested that a new framework is needed for such modeling (Brynjolfsson and Mitchell 2017), the results of this study indicate that the current frameworks being explored are also inadequate. Judgmental techniques have been used in earlier work to forecast levels of labor displacement that are nearly to the modestly extreme level assessed here (Frey and Osborne 2017), but no techniques beyond expert surveys exist modeling or forecasting the more extreme levels of labor displacement that experts working on discontinuous AI technologies give a 10% chance of occurring within a decade.

These implications for future of work researchers are significant because the consequences of making errors regarding such extreme labor displacing AI scenarios are severe. While this is not the focus of this study it underscores the significance of examining techniques for forecasting AI progress more closely and open mindedly. While the proposed holistic forecasting framework is intended for the purpose of forecasting technical AI progress toward different levels of TAI, the notion of a productivity bonus that is tied to DTAI was not considered in much detail in the framework or in the JDM model. Thus, tying together forecasts of the type expected to be produced

[144] Forecasts by these experts have significantly less uncertainty than those from other AI researchers.

by JDM with forecasts of labor displacement is something that is highly desirable. In the holistic framework this could be considered to be an implicit result of the strategic planning portion of the framework, however, this may not be a sufficient solution. Or, in other words, perhaps a separate model should extend the results of JDM by an additional round of interviews or questionnaires with economists or specialized AI/robotics researchers in order to best model the labor displacing effects of TAI/DTAI/RTAI. This is certainly something that should be considered for future work.

Surveys have long been the primary means for forecasting transformative AI, albeit the primary interest of previous studies has been RTAI (Baum et al. 2012, Muller and Bostrom 2016, Grace et al. 2018, Zhang and Dafoe 2019). There is no reason to expect this trend to end, but this study, by exploring different levels of extreme labor displacement from AI, has introduced a novel and perhaps more useful approach for forecasting severe impacts of AI. Because previous studies focused on notions of AGI and the automation of specific technologies, they were limited in their ability to capture respondents' perceptions of larger labor displacement trends that may occur between now and the development of RTAI. Or, more generally the previous studies were limited in that they did not ask questions about TAI/DTAI/RTAI independently.

The results also have implications for the techniques for analysis and aggregation of results from future surveys concerning RTAI. The fact that this study modeled both median forecasts and uncertainty is a unique feature which can be very significant for future survey analysis. For example, if Grace et al. (2018) were to conduct the same survey again, with the same participants, modeling of uncertainty could be very valuable. Specifically, if the original data and new data were modeled with a dummy variable, and this dummy variable is statistically significant with a negative beta value this would indicate that experts' uncertainty was decreasing with time. This would suggest that the most recent forecasts are more actionable, due to the decreasing uncertainty.

Particularly, if the forecasts were decreasing, this would suggest the need for quicker action to mitigate risks.

While not successful, the naive calibration technique that was utilized is a step in the right direction for controlling overconfidence. This technique could be explored further in future work to see if it, or modifications of it, could be made useful to survey administrators. Despite its failure, the inclusion of a method for controlling overconfidence is significant, and something that future studies could benefit from. Another alternative would be to use calibration training for participants. While online calibration training of as little as one hour has been demonstrated to be successful (Moore et al. 2017), it may be possible to introduce a quick 10 or 20 question module that could be completed in roughly the same time as the calibration questions which could help to control for overconfidence (Gruetzemacher et al. 2020).

8.1.2 Implications for Forecasting Transformative AI

In general, the results of the survey portion of this study seem effective at reducing uncertainty, and this is the ultimate goal of forecasting; to reduce uncertainty for decision makers. Moreover, for DTAI and RTAI it may be the best we can do – this has certainly been the case in the past. However, the other methods utilized in this study, while not validated or evaluated for accuracy, do appear promising for reducing uncertainty much more effectively than the survey.

Unlike statistical models or extrapolative forecasts, the forecasts generated from surveys are not well suited for expert adjustment or – given the experience from this study – for incorporation in a holistic forecasting framework. Based on the results from all methods demonstrated in this study surveys seem appropriate only when there are limited resources or when only one single method can be used. Furthermore, the collective results of this study suggest that surveys may be best suited for generating forecasts that can be distributed publicly due to risks from information hazards (i.e., risks from the dissemination of information that could be hazardous to humanity if obtained by irresponsible actors; Bostrom 2011).

8.2 Tech Mining

While there were no minor components of this study, the demonstration of tech mining was not given as much focus as other methods. This was not because the method was perceived to be inferior, rather, the method is widely thought to be one of the most effective techniques for forecasting AI progress (Gruetzemacher et al. 2020). This assumption is based on the prevalence of tech mining for different types of technology forecasting[145] (Das et al. 2002, Martinez-Plumed et al. 2020). However, the demonstration here still showed, if only qualitatively, the potential value of tech mining for technological forecasting.

While results of the clustering were not very strong for any of the datasets, the results from the NIPS proceedings were better suited for forecasting. This was demonstrated by looking at the development of the clusters generated from the data over time. Qualitatively, it is clear that substantial progress in certain clusters occurred over different periods. The magnitude of clusters could be measured and extrapolated over time, or the change in magnitudes of different clusters could be used as independent variables for a more complex model. Furthermore, only an n of 10 was considered and minimal effort was made to tweak hyperparameters or to explore the data in depth due to the time intensive nature of such work. Because searching for signal in text data is such a challenging task, for demonstrative purposes the tech mining portion of this study is sufficient.

Tech mining is very well suited for judgmental adjustment, similar to forecasts from statistical models or from extrapolation. This is because tech mining can be used to identify a large

[145] Also considering the large-scale think tank effort.

number of indicators (Porter and Cunningham 2004), which can also be extrapolated or used for statistical modeling. While quantitative forecasts were not generated using the tech mining demonstration in this study, it was shown that time dependent trends could be identified from the data when the data was rich enough. The demonstration here was rich enough for this when using full-text manuscripts/papers. However, datasets orders of magnitude larger could also provide a rich enough dataset for identifying such trends with the potential for modeling or extrapolation.

8.3 Indicator Extrapolation

Only two of the forecasting methods examined in this study actually yielded an actionable forecast (as designed): the survey and the indicator extrapolation. The forecast from the survey could be thought of as a forecast of DTAI while the forecast from the DeepMind operating costs extrapolation could be thought of as a forecast for RTAI.

The forecast generated from extrapolating DeepMind's operating costs is a strong forecast because the economics of previous major scientific achievements suggest that operating costs would be a good indicator of progress toward a grand scientific objective. The inclusion of the levels of funding associated with previous major scientific achievements is included to illustrate the timelines that could be expected with the increasing resources that Google is directing toward DeepMind. Considering these up to levels that could enable DeepMind to reach some level of RTAI suggests that such powerful AI systems could be alarmingly close. Moreover, based on projected income Google could support even the most extreme levels enabling it to continue increasing DeepMind's funding through 2026 to levels associated only with a project on the scale of the Apollo program.

A third actionable forecast could be considered to be the aggregated expert adjusted forecasts (Sanders and Ritzman 2001) of the indicator extrapolation. These are not reported, but the participants in the 2nd round of the JDM interviews were asked to respond to the DeepMind operating costs extrapolation. We choose not to report these because we do not place any value on the individual forecasts and assigning expert opinion to this extrapolation could give it more significance than it is due. For this reason, we believe that such forecasts could be considered information hazards (Bostrom 2011).

Despite interviewees being asked about this forecast during the second round of the JDM interviews, this forecast is not a good component forecast for JDM or for the holistic forecasting framework proposed here. It, like the practitioner survey, would be most useful as a standalone forecast or as a data point for decision makers to consider. Moreover, it is something that is easily monitored, much more so than an indicator like the computational resources required to achieve major AI benchmarks, like that proposed by Amodei and Hernandez (2018). Their indicator may appear to be easily measured, however, the computational resources required to train very powerful systems are not published with the results of studies, and there exists few incentives for research organizations to publish such results, likely because the publication of such results may suggest that the research is more reliant on the massive resources available to only a few companies and less a result of research progress. This image could hurt future talent acquisition, yet, considering multiple dimensions in analysis of forecasts and for building forecasting models is something that should receive more attention (Martinez-Plumed 2018).

8.4 Mixed Methods Analysis

The results from the mixed methods analysis were limited because the sample size of interviewees was small and because it was difficult to elicit probabilities for each of the survey questions for the majority of those who did provide probabilities for the survey questions. The fact that two participants were not comfortable making quantitative forecasts for their area of expertise underscores the difficulty of finding experts who are also able to give quality estimates of future progress. These two participants who did not provide numeric responses to the forecasting questions expressed their lack of expertise regarding forecasting and the intractability of forecasting AI progress as reasons for this. It is interesting that two of twelve randomly sampled AI experts who did agree to participate in an interview explicitly on the topic of forecasting AI progress still held these opinions[146].

8.4.1 Quantitative Comparison with Survey

Because the responses from interviewees indicated that they were being very deliberate when making their forecasts, and because many of the interviews involved long pauses or participants vocally talking through their thought processes while arriving at their forecasts, the quality of the resulting forecasts seemed to be the results of more focused thought than the surveys. This hypothesis was formed from the perceived amount of time spent on answering the survey questions during interviews compared to the survey results dashboard indicating a mean survey completion time of 13 minutes.

To test the hypothesis that there was no different between the amount of time spent by interviewees completing survey questions and the survey respondents completing survey questions, the time taken for each was collected and a two-sample t-test for equivalence was conducted. The result was that there was strong statistical evidence to reject the hypothesis and conclude that there was in fact a significant difference between the amount of time spent completing interview questions by survey participants and interviewees. This can be interpreted in

[146] One of these experts expressed his inability to give quantitative answers and his skepticism for the plausibility of forecasting AI progress before consenting to participate. It was decided that despite these caveats, letting him continue to participate would yield a sample better representative of the population.

many ways, and the experiment was not designed with enough control to support any of these interpretations, but they are still discussed here.

One significant way to interpret these results is that many of the survey participants responded with system 1 responses while the interviewees, with one exception, appeared to all give system 2 responses (Kahneman 2011). This is almost certainly an oversimplification because the terminology and definitions used were intended to avoid existing cognitive biases based on familiarity of terms, but it is possible that many responses to the survey were entered based on an initial reaction without any more analytical thinking. Just under half of all responses under an hour were below 13 minutes (67/136), with 40 under 10 minutes, 17 under 8 minutes and 6 under 6 minutes. Given the amount of content, and the fact that quickest time to complete the survey questions in an interview was 15.5 minutes, it appears that a significant portion of the respondents likely responded without thorough consideration. It is important to note that the survey participants were all practitioners, so this could have played a role as well, however, it does suggest there is substantial value in interactive elicitations when the value of the data is enhanced.

Because the interviewees were all experts, as opposed to practitioners who participated in the survey, it was difficult to determine whether levels of confidence, overconfidence or uncertainty were at levels that could merit weighting experts differently. However, this is certainly something that should be explored more in future work. Based on the results, it does appear that rigorous, well designed qualitative experiments could validate methods for determining weights based on behavioral signals exhibited during interviews.

A very interesting research area would be to determine if deep learning could be used to extract emotional or behavioral features from video during interviews that could be used to predict forecast accuracy or precision. This would be straightforward to do if a pretrained model for emotion recognition was trained on a forecasting dataset. The primary challenge here would be to create a dataset of videos capturing participants reactions while responding to forecasting questions. This could perhaps be done without using a human facilitator, through an app, perhaps a calibration training app. It could also be done using a video chat. The performance would have to be incentivized as well as the participation itself to ensure that participants make earnest efforts to make their best possible forecasts.

Overall, this test suggests that participants do not spend as much time thinking about forecasts when making the forecasts through an online portal. No conclusions can be made about the quality of such forecasts without a further experiment to evaluate the accuracy of forecasts made during online surveys compared to interactive elicitations. This is also a very strong possible avenue for future research. Until such study has been conducted, conclusively determining one way or the other if there is an impact on forecasting accuracy from the different elicitation formats, interactive elicitations should be used by default. This conclusion is based entirely on the fact that more time is spent, and Kahneman's work suggests decision making is improved under such conditions (Kahneman 2011).

8.4.2 Qualitative Elements

Regarding the qualitative observations, much of the important implications were discussed in the previous section. In general, there were a lot fewer complaints about questions and language, likely because clarification questions were possible because of the interactive nature of the elicitation. Thus, one clear advantage of interactive elicitations is the significance of semantic ambiguity in definitions of key terms or in the phasing of the questions themselves.

Regarding the qualitative questions, it was interesting to see the diversity of perceptions among experts of the five years of progress in AI that had preceded the interviews. There appeared a slight bias toward the significance of recent progress by younger interviewees and a slight bias against the significance of recent progress by the experts who had been in the field the longest. Furthermore, those who were mid-career academics, or less than incredibly accomplished, perceived the rate of progress in the field of AI as having increased slightly in recent years due to increased investment and talent, but not proportional to this increased investment and talent.

Regarding indicators of true progress, it was interesting to observe the reactions of interviewees which signaled that they thought true progress diverged significantly from what was commonly perceived as progress. Moreover, it was interesting that the three most common themes of responses to this query are now likely the three most promising research directions (two are very promising: robotics and NLU; the other is the most highly debated and the one expected to have the greatest long term impact on AI capabilities: combining symbolic reasoning and deep learning (Marcus and Davis 2020)). This appears to indicate that well selected experts have foresight that is effective 18-24 months out. This would be consistent with Tetlock's observations on the typical forecastable window (Tetlock and Gardner 2016). Also, consistent with Tetlock's conclusions, it is possible that some forecasts are valid for longer periods, but there is no evidence to support this other than the fact that self-driving vehicles seem to be perceived as further away now than they were at the time of the interviews. Given all mentioned here and reported in the discussion, it appears that qualitative forecasts from experts are valuable in aggregate, but that individual expert forecasts, even from some those who have done some of the most innovative work in a particular field can be very misleading or difficult to interpret[147].

[147] One of the most accomplished experts interviewed appeared to take the notion of *true* progress to indicate progress toward human level intelligence. Thus, regarding future progress in his particular area of virtually unparalleled expertise, he was very dismissive of his predictions for the next five years although these predictions are being realized and, among AI experts, are thought to have significant transformative potential for labor markets. His opinions and further insights on these predictions was not elicited in follow-up questions because in the course of the conversation his demeanor suggested that the predictions were more trivial than they now appear to be. To be clear, there was some

The final qualitative question posed the greatest challenge to the interviewees, and simply interpreted, seems to indicate the extreme difficulty in forecasting future progress toward DTAI and RTAI. The primary suggestions that were given, such as theory of mind and metacognition, could be interpreted to indicate that the lack of our own understanding of intelligence is a major hurdle in being able to forecast the development of DTAI or RTAI. For things like nuclear power, and nuclear fusion in particular, it has been clear that this is a possible source for energy and that the bottleneck is not sustaining a fusion reaction, because existing fusion reactors can do this. Research on fusion reactors has been ongoing since the 1940s, and fusion reactors have been demonstrated for nearly as long; there has yet to be a fusion reactor with positive efficiency because the problem of confining a self-sustaining fusion reaction requires a large amount of energy. Furthermore, a means for extracting net positive energy if produced is also a challenge. In other words, the challenges for commercialization of nuclear fusion reactors are engineering ones and not fundamental science. In the case of AI, the foreseeable challenges appear to be more fundamental. This would suggest greater risks from a breakthrough because, if the engineering is as simple as many anticipate, then safe engineering practices may be overlooked or ignored. Moreover, engineering safety is not something that computer science has a strong record with. However, on the contrary it is also possible that even if the fundamental challenges were overcome the required engineering may be much more challenging than anticipated and there may be ample time to develop safety mechanisms once more is known about the engineering problem that is the bottleneck.

ambiguity in his response, and it is difficult to know what the true intentions were for certain. Based on all of the relevant information it does appear his predictions were very prescient, which would be expected given his expertise and the relevance of his groundbreaking work and foresight to recent progress.

8.5 Scenario Network Mapping

The SNM workshops were very useful for identifying strengths and weaknesses of the SNM technique for the purpose of mapping plausible paths of technological development of AGI as well as for mapping the plausible scenarios for the development of AGI. Moreover, because there was so much variance in the three different groups that participated in workshops, the workshops were useful in shedding light on how the technique might also be used for strategic planning given an input scenario map (e.g., from JDM). Some suggestions for future use and development of SNM for both of these purposes are described below.

The first workshop had the best outcome regarding the plausible paths to AGI, and this can be attributed directly to the participants' knowledge of the material and their motivation for creating the map. Alternately, the failure of the second workshop to create a map is also a result of the participants' knowledge of the appropriate subject matter and their motivation. One significant factor in the success of the first workshop may be in the diversity of expertise of the participants. While all participants were knowledgeable regarding different aspects of AGI development, there are numerous different ways to develop AGI. Specifically, there are paths that are more neurologically inspired, cognitively inspired, algorithmically inspired, logically inspired or agent inspired[148]. Despite having diversity in the group of participants in the first workshop, this diversity is only significant relative to the diversity of technical expertise from either of the two other groups that participated in SNM workshops. Relative to the diversity of the field, or with respect to the

[148] This is a quick and woefully insufficient categorization of the different plausible paths to AGI. An updated paper on this topic, similar to Goertzel (2014a), from the perspective of someone not actively involved in the research of one of these or other possible categories of AGI is something that could be useful. The example here is simply included to demonstrate the variety of possible routes.

different paths provided as an example, the expertise from this first workshop was relatively narrow[149].

List (2005) and Gaziulusoy (2010) were both very clear in their descriptions of the significance of using a diverse group of experts in SNM. Thus, for this study efforts were made to incorporate as wide a diversity of experts as possible, however, this was just not realistic for developmental workshops. That three workshops were able to be held using participants of differing levels of expertise regarding AGI[150] was a major success. Moreover, the three different groups offer significant insight into how effective the technique is and what mechanics of the workshop process are the most ripe for improvement. Generally, the AGI research community does not focus as much on agent inspired AGI architectures as on biologically inspired AGI architectures. The participants of the first workshop included participants from each of these, and this diversity was seen to be a major constructive element leading to the positive outcome. In the second workshop, one participant had substantial machine learning experience, but this participant was able to participate in one of the two groups that happened to be the group that was substantially more productive. The participants of the third workshop lacked substantial technical expertise or expertise on AI strategy issues, thus, this resulted in a map that was not very granular. However, the resulting map was useful for high-level strategic planning purposes and is an indicator that the technique is useful for groups of non-experts or decision makers (as would be consistent with List and Gaziulusoy)[151].

[149] The diversity of different paths to AGI is also evidenced by the numerous different approaches to AI safety (Everitt et al. 2018). There are at least four different published research agendas on the topic (Lieke et al. 2018, Soares and Fallenstein 2017, Amodei et al. 2016, Taylor et al. 2016).

[150] AGI expertise is a particularly esoteric area of expertise making expert participation especially challenging.

[151] It should be noted that while the technical expertise was lacking, the diversity of expertise ranged from futures studies to psychology to machine learning to existential and catastrophic risks. Thus, the diversity was acceptable.

Another issue that is crucial to the success of an SNM workshop is the proper venue and space. We were fortunate during the first workshop to have an ideal sized room that allowed for pushing two tables together for group work and for breaking apart for subgroup work. Further, the room had a large, curved wall that offered an adequate space for pasting the working scenario network map onto. This room was also located in the retreat hotel and was available for participants to visit at any time between the two different workshop sessions to reflect on the results from the first session, which seemed useful. The space for the second workshop did not provide a wall whatsoever for posting the working scenario network map to, and this was certainly a problem. Moreover, this problem should not be understated; while lack of diversity and participant motivation was found to be the primary cause of failure, the lack of a proper space was likely the reason that some final map was not generated. Had there been a proper space with the map mounted on the wall, a map would have likely been generated although it would still have been incomplete, and some differences may have been difficult to resolve.

One difference between the first workshop and each of the remaining workshops was the number of participants[152]. The first workshop had a small number of participants which made the group working sessions much easier to manage. The inclusion of more participants had several effects that were challenging. For one, the increase in the number of participants seemed to have a direct effect on the amount of time required for generating maps. The first workshop was split into two sessions; in the first session a historical scenario network map was generated and in the second session a future scenario network map was generated. For the second workshop two sessions were unable to successfully generate a single future scenario network map. For the third workshop, four half-day sessions were required to generate a single map. Although a significant

[152] The first workshop had five participants, the second workshop had eleven participants and the third workshop had an average of nine participants per day.

amount of this time can be attributed to the use of the NeyroKod software, and to a more thorough workshop structure, based on the experience the generation of one map would likely require two workshop sessions.

List suggests using assistant facilitators when there is significant disagreement or a large number of participants (2006). Moreover, List suggests that the ideal size of workshops would involve 15 to 20 participants. Based on the experience here, using the technique for the task of mapping an actual technology's development, this suggestion does not seem appropriate. It is likely that there is implicitly significant disagreement because of the highly uncertain nature of RTAI technologies in general, and further disagreement given the fact that there is really only one proven means for developing AGI which is fundamentally impossible (Hutter 2005; i.e., there is no consensus regarding the effective ways to develop AGI). The third workshop did utilize co-facilitators, and this was beneficial, especially given the challenges of this workshop due to the incorporation of the NeyroKod software. The second workshop did not use co-facilitators, which put a substantial burden on the single facilitator. The use of a co-facilitator was planned for the second workshop, however, the second facilitator chose not to participate in that role during the workshop.

Another major difference between the first and third workshops and the second workshop was the amount of influence exerted by the lead facilitator. Drawing on expertise from previous interviews, the lead facilitator used a heavy hand to guide the process for both the first and third workshops. This influence served only to keep the resulting map within the scope of the most probable paths to AGI rather than to have a granular influence on the resulting scenario network map. For the second workshop, however, given significant conflict between parties the facilitator opted to let the participants have complete control over the outcome by not influencing the map at all. Given the other factors negatively impacting this workshop, this was ineffective. This technique may be more effective with higher quality experts but based on the previous attempt this is also unlikely (Goertzel 2014b, Goertzel 2016). Thus, strong and rigid facilitation is suggested for generating usable technical paths to AGI. One option would be to use a basic map to begin, perhaps similar to a very crude scenario map like those generated from JDM. Then, more granular details could be explored by the participants.

The experience gained by facilitating these three developmental SNM workshops was very valuable for confirming the utility of the technique and for identifying things to improve upon. If an organization was interested in utilizing SNM for the generation of the most plausible paths to AGI now, it would be suggested that the organization plan for a two part workshop, most similar to that of the first workshop, with the option to extend the workshop to a third day if necessary to unroll some things only identified during the event tree generation stage. Moreover, it would be suggested that a maximum of ten participants were involved, and that these participants remained constant for all workshops. (Other recommendations consistent with the previous commentary in this section would also be made.) However, there is much room for improvement of the methods, and this is discussed more in Chapter 10.

Previous discussion in this section has focused on the success of the SNM workshops for developing technical maps of the plausible paths to AGI (i.e., containing the component technologies). However, in the second and third workshop, more elements of scenarios surrounding the development of AGI were considered. Time ran out in the third workshop, but plans were made to include a separate section for creating a separate milestone layer. In general, different layers could prove very useful for numerous objectives. Moreover, different layers could be added to the same map from different groups that represented different areas of expertise. For example, a group of technical experts could use the workshopping process to generate a technical map of the plausible paths to AGI, and a group of AI strategy experts could participate in a separate SNM workshopping process to extend the map from the first workshop to model different coordination challenges for stakeholders attempting to develop a safe AGI.

While the second workshop was ultimately a failure, the third workshop did demonstrate that it was possible to use the process for the task of developing maps that were useful for strategic planning, as is suggested in the holistic forecasting framework proposed in this study. It is likely that this would be best done using either an earlier SNM workshop for identifying the technical paths or a JDM process. However, the primary benefit of JDM in scenario mapping is that it can offer quantitative, probabilistic forecasts to assist decision makers. If this is not desirable or is for some reason irrelevant, then the use of SNM alone may be a suitable or even preferable route for strategic planning.

In general, for both purposes (i.e. technical mapping and strategic planning), the use of NeyroKod seemed to be a significant advantage for the process. However, because involving its use in the workshopping process diverges so significantly from the instructions for the standard implementation of SNM (List 2005), substantial work remains before such a version of the workshop can be made operational. This is likely where future efforts should focus regarding developmental workshops, as the two to three session workshop using the tactile process would likely be sufficient for existing purposes.

8.6 Judgmental Distillation Mapping

In general, this study promotes the use of the class of methods identified as scenario mapping methods to be well suited for use for the purpose of forecasting TAI/DTAI/RTAI. The demonstration of JDM was one of the most significant components of this study because it represents the scenario mapping technique that is promoted for the purposes of both qualitative and quantitative forecasting. Moreover, the technique is central to the holistic forecasting framework proposed in the study.

For demonstration purposes only the qualitative aspects of the JDM process were conducted because demonstration of the full JDM process to generate quantitative forecasts was beyond the scope of this study. For qualitative purposes the demonstration was successful in generating a scenario map of the paths to RTAI that included weights for the different causal relationships of the external nodes. This was done through three rounds of interviews with the first round focusing on generating a basic map, the second round focusing on refining the map and the third round focusing on identifying the weights as well as demonstrating the decomposition of one of the external nodes. Substantial insight was gained from this demonstration of the JDM process which can be useful for those who want to use the technique to generate forecasts or those who would like to develop the technique further or incorporate elements of it in other expert elicitation techniques.

Because empirical distributions were not elicited for the external nodes through forecasting techniques, little can be said about the actual mechanics proposed for JDM. While the demonstration of the Monte Carlo simulation through parameterized distributions was effective for the purposes of this study, future work should certainly evaluate the proposed approach for combining forecasts. Numerous parameters, such as discontinuous progress, could be considered increasing the complexity of possible models and making it more challenging to identify the best model.

8.6.1 Iterative Expert Elicitation

Because this was a novel technique and there was no prior experience to draw from for creating the scenario map using questionnaires and interviews, the iterative expert elicitation was designed with little information regarding the quality and volume of data needed for each round. Thus, much was learned about how these rounds can be best designed and implemented through the experience of demonstrating the technique for this study. Specifically, the first round of interviews was effective, and if anything, could be sped up. It could even be possible to combine the basic map generation round and the refinement of the map round into a single round of interviews. This would be the case if the round of interviews was conducted over a longer period of time, e.g., multiple conferences, otherwise distinct rounds would be necessary. Thus, perhaps there should not be a rigid designation of only three rounds of questionnaires/interviews for the iterative expert elicitation process in JDM.

Another insight about the design of the iterative expert elicitation was that questionnaires would likely be more useful for some elements of the process than interviews. Interviews were best suited for this study because it would have been more difficult to get AI experts to give the same amount of time to the elicitation if questionnaires had been used. This challenge would be more easily overcome if a well-respected organization, government or academic institution was leading the effort (even more so if experts were in the lead investigator’s network or if the lead investigator had name recognition). Interviews were challenging for numerous reasons. For one, they had to be conducted close together at conferences in a limited amount of time. Moreover, it was typically unknown to the investigator who would choose to participate in an interview before the conferences began, and much effort was spent at conferences attending talks or tailoring email invites in attempts to solicit respected experts on specific topics. Consequently, it was difficult to create specific interview questions for each expert (it was difficult to prepare for more than one hour for each expert because of the fast-paced environment of conferences) without sacrificing quantity of interviews.

The challenge of balancing quantity and quality of interviews was something that was handled in different ways for the different rounds of the process. The first round of interviews focused on trying to balance quantity and quality. While this technique was successful at identifying a generic scenario map, it was not very successful for digging deeper into the more granular details of the different paths toward the development of DTAI and RTAI. Moreover, the method of expert selection was similar to that of the mixed methods interview where any highly esteemed AI experts' opinion was sought after. Consequently, the second round of interview focused more on quality as opposed to quantity. This was achieved by more targeted expert selection, focusing on experts with expertise in areas directly related to the most prevalent topics identified in the earlier round of interviews. Consequently, less experts were approached, and it was observed that directly approaching experts following presentations or during poster presentations led to higher participation rates for interviews, so this was used in lieu of (or often in combination with) email solicitation. Efforts were also made to conduct fewer interviews such as to focus on the quality of experts as well as the quality of questions and preparation for the interviews. This appears to have been successful because several of the interviews from this second round were crucial to the refinement process of the final scenario map generated.

Another technique was attempted in the third round, because the needs for this round again changed. The third round required a larger amount of specialized expertise, but not significant amounts of time or input from each expert. Thus, experts were approached informally and asked more general and less direct questions regarding the information and expertise attempting to be elicited. This involved approaching experts working directly on research tied to the nodes in the existing JDM and asking questions about how this research related to other nodes in the JDM or how the research might be broken up into sub problems. Because so much information was required of this round, even through questioning 17 experts in this manner only the minimal information for the demonstration was collected.

Based on the experiences outlined in the previous paragraphs regarding the design and implementation of the iterative expert elicitation stage of the JDM process, several suggestions can be made for future work. For one, the iterative elicitation process should be flexible rather than rigid, and this is more easily done without having to approach experts directly without any introductions. Thus, in a well-connected and well-funded organization experts should be approached and solicited for participation with generous compensation for their time (e.g., at least similar to what their standard compensation is from their employer). Furthermore, questionnaires would likely be more useful all rounds, especially if their use allows for more time customizing the questions and if their use allows for staggering their dissemination such that data is collected on a rolling basis and analysis and map development can be a continual process. This would enable adequate time for analysis of previous experts' responses prior to finalizing questions for the next experts' questionnaires. Moreover, it would allow for interactive modification of the map or for inclusion of detailed instructions regarding specific components of the working map. Furthermore, it would allow for revisiting experts to identify weights or for decomposition questions – tasks requiring much less time from experts – in the later rounds of the iterative expert elicitation process. This should be explored for future work, and it may be possible to explore this possibility through mapping a technology other than an AI technology in order to refine the process while not turning off potential future experts by imposing too extreme of a demand on their time (if possible this would enable these suggestions to be explored without having to compensate AI experts at the exorbitant rates that the leaders in the field are currently paid[153]).

8.6.2 Models for Propagating Forecasts in Scenario Maps

Monte Carlo simulation was proposed as the method to be used for propagating forecasts in the scenario maps generated by JDM. Bayes nets might seem like the first choice for this task, however, they are poorly equipped for handling empirical distributions that would result from the elicitation techniques used to generate the forecasts for the external nodes and their decomposed elements. Monte Carlo simulation, on the other hand, is well-suited for handling the empirical distributions because it simply involves random sampling. Thus, if forecasts are to be elicited as empirical distributions, Monte Carlo simulation is preferable to Bayes nets. For most cases, it is significantly more desirable to operationalize the information obtained from experts' forecasts in the form of an empirical distribution, so Monte Carlo simulation seems likely to be a topic for future work to explore if JDM is to be used in practice.

There are different ways to handle Monte Carlo simulation when using empirical distributions as is the case for the scenario maps generated by JDM. For the demonstration shown in Chapter 5, weights were normalized for aggregating the results of the sampled distributions and discontinuities were modeled using stochasticity to determine whether an input node was sufficient for discontinuous progress. Many different variations of this exist, and it is worthwhile to explore these in more detail. Based on how discontinuities are handled and how weights are used for aggregation, the resulting forecasts can be very different, even with the same input distributions. This should certainly be explored further because this study did little in this regard.

[153] Other experts of the same class from different domains may be paid three to five times less for equivalent levels of compensation as many of the most desired AI experts are earning annual salaries well above $500,000 (new hires are commonly paid between $300,000 and $500,000 annually).

One final desirable element for the Monte Carlo simulation, and more generally, for the quantitative forecasts generated using the JDM technique, is evaluation and validation. If the technique could be validated, it would then be a real option for forecasters. Until then, it is a solid theory, but still unproven for practical application. For the purpose of validation, a simple toy problem should be identified that has multiple external nodes and a single next generation technology layer (ideally one that can be resolved in the near future involving technologies that are near-term plausible). The external nodes should also be decomposable into tractably forecastable targets. Even demonstrating JDM's validity for a toy problem may be challenging, but, if successful, it would certainly be worth the effort because no existing techniques are able to combine qualitative and quantitative elements in the same way. Thus, if validated, the technique would be incredibly useful for potentially a larger variety of practical applications in technology forecasting or the forecasting of other complex events within scenario maps.

8.7 Holistic Forecasting Framework

The holistic forecasting framework is yet another crucial element of this study, and, in combination, the majority of the different techniques demonstrated in this study were effective as a proof of concept for the specific holistic forecasting framework proposed here (see Figure 9). This specific holistic forecasting framework primarily focused on the JDM process, which was discussed in the previous subsection. Because the JDM process demonstrated here was limited and was not used to generate a quantitative forecast for the entire map, including the end states, only the framework's techniques for informing the JDM process were demonstrated. Particularly, the use of information gained from earlier forecasts for informing the input of the JDM process, including the practitioner survey, the tech mining, the indicator extrapolation and the mixed methods analysis were used. While not a part of this study, the use of wargames or role-play games for strategic planning has also been explored and is discussed in this section.

The practitioner survey, tech mining, indicator extrapolation and the mixed methods analysis were all components of this study that were conducted prior to the completion of the JDM process (see Figure 5), and, based on the proposed holistic forecasting framework, were well suited for informing the JDM process. Because these elements were completed prior to JDM, it is hard to state that any of them had no influence on the JDM process, however, some of them appear to have had more significant influence than others; the mixed methods interviews seem to have had a particularly significant influence on the JDM interviews. While these interviews were structured, a small number of questions were included that elicited qualitative data from experts that the remaining questions, which were also included in the survey, did not. These questions were very effective at identifying topics of interest for focus in the JDM interviews. The indicator extrapolation and the tech mining demonstrated here were not very useful for informing the JDM process, nor was the survey.

While the mixed methods interviews seemed to be the most useful input forecasts of the JDM process that is likely because some interview questions allowed for great flexibility to explore experts' expectations of future AI progress. The other techniques that were input candidates for JDM were not tailored for this purpose. If indicator extrapolation, tech mining and the practitioner survey had been conducted with the intention of being components of the larger process, they may have been more useful. However, they were conducted for the purpose of demonstrating and evaluating their independent effectiveness as forecasting techniques for forecasting TAI. Thus, they did not validate the use of multiple techniques in a holistic framework in a very robust way.

The first SNM development workshop was completed shortly after the end of the first round of JDM interviews and shortly before the beginning of the second round of JDM interviews. For this reason, it was not incorporated into the first round JDM interviews, however, information gained from this workshop was the most useful for informing the later rounds of the JDM interviews. Specifically, some of the information gained in this workshop was directly incorporated into the map.

Generally, informing the JDM process with some sort of input mapping or forecasts is likely unnecessary. However, using other techniques for informing the process of generating the qualitative map can be very useful, and can effectively improve the quality of the resulting map. Particularly, powerful qualitative forecasting techniques, such as other scenario mapping methods like SNM, can be very useful for enhancing the opinions elicited directly from experts in the iterative rounds of interviews. A major challenge of the JDM iterative expert elicitation process is the aggregation of different types of questions for input regarding different technologies from experts with different types of expertise. One way to possibly deal with this is to use results from early rounds of JDM as an input for an SNM workshop, in which these results could be aggregated with group consensus during the workshop and then the resulting map could be condensed for further rounds of the JDM process. Combining these two very powerful scenario mapping techniques may be the best way for leveraging expert elicitation regarding qualitatively understanding the likely trajectories of future AI progress.

The strategic planning element of the holistic forecasting framework proposed for this study has been largely ignored as it was beyond the scope of the study. The use of SNM was suggested for adding layers to the scenario map resulting from JDM or to a scenario map resulting from an earlier more technical SNM workshop, however, this was the only reference to this element of the holistic forecasting framework. The output scenario map from JDM would likely be well suited for use with more common scenario analysis techniques such as those of the intuitive logics school, which would be a useful form of more technical forecasts for distributing to policy makers and other decision makers (Roper et al. 2011).

Another useful technique would be the use of simulation or serious role-play games for training policy makers and decision makers (Avin et al. 2020). This, like scenario analysis using the intuitive logics scenarios, would make some of the more dramatic or radical societal transformations possible from future progress in AI more palatable to the target audience through an experiential learning process. This technique can also be useful for exploring plausible futures and can create a feedback loop for informing future rounds of SNM or intuitive logics scenario planning. A similar feedback loop is possible from the JDM process, for improving the resulting map. Even the FICT notion proposed by Tetlock (2017) includes a similar feedback loop. Because the feedback loop is a feature of holistic forecasting frameworks broadly, the strategic planning element of the holistic forecasting framework proposed here could be thought of as an independent forecasting framework. Future work could explore developing a more specific framework for this strategic planning element.

Another major element of the holistic forecasting framework is that the value of the framework is not limited to the specific framework proposed here. The idea that more holistic frameworks should be used for forecasting complex technologies or political events is a novel idea itself. The framework proposed is one of but two techniques that meet the criteria for holistic frameworks stated here; the other[154] is the full-inference-cycle tournament (FICT; just mentioned) technique proposed by Tetlock in an Intelligence Advanced Research Planning Activity (IARPA)

[154] As previously discussed, it is possible that the extensive AI forecasting work by a Washington DC think tank now constitutes a holistic forecasting framework.

grant proposal (Tetlock 2017). Based on the results of this study, some suggestions can be made for yet another framework that meets the criteria stated here, but this builds on the results from the Delphi technique. Thus, this is discussed in the next subsection.

8.8 The Delphi

The Delphi technique was the only technique that was demonstrated for something other than forecasting (i.e., for creating a research agenda). However, there were two reasons for applying it for this alternate purpose: 1) because a research agenda was a needed item for those interested forecasting AI progress, and thus a significant contribution to the existing body of knowledge, but also 2) because the adapted Delphi process develop has great potential to be incorporated into Tetlock's FICT proposal for the purpose on including expert opinions in full-inference-cycle tournaments.

First, regarding the generation of a research agenda: this was a success. Many different opinions are present regarding the role of forecasting AI progress among AI strategy and AI safety researchers, however, there is consensus on the significant role that AI timelines play with respect to peoples' own research and career plans for working in these areas. These many different opinions can be particularly challenging for new people in these research areas, as the information sources related to AI forecasting that are available online are not necessarily very representative of the consensus of academic and professional work on the topic. Thus, there was a need for a unifying document, that contained verifiably objective information regarding what are the most important topics to consider for forecasting AI progress. This could help keep members of the AI strategy and AI safety research communities from spending time and resources studying forecasting work that may not be relevant to that of the experts' consensus. This can also help such people to determine for themselves how to weight different timelines and different predictions so that they can be better informed. Moreover, it can help direct such people's thinking so that they may offer constructive insight into the methods and techniques that could possibly be used for forecasting AI progress. For these goals, this effort was successful, and the result will be published soon (Gruetzemacher et al. 2020).

Second, it is very important to continue to pursue the development of different forms of holistic forecasting frameworks. The use of the adapted policy Delphi process used in this study effectively demonstrated the ability of this technique for generating a ranked list of different types of questions relevant to a specific topic using true expert opinion. While the technique had several limitations (to be discussed in Chapter 10), when addressed properly, future iterations will certainly be very effective for this task. Thus, the technique demonstrated here would be suitable for incorporating expert judgement into existing techniques, such as Tetlock's (2017) full-inference-cycle tournament notion. It could also be used in other ways to incorporate expert judgement for forecasting target generation that could then be combined with other forecasting techniques to create alternative formulations that meet the criteria proposed here for holistic forecasting frameworks.

If the technique were to be used in alternative future versions of a holistic forecasting framework, it may be useful to spend more time developing a platform for facilitating the Delphi process used here. This would be particularly important with large numbers of experts, or if there were a large number of topics as there were for the different questions of interest for forecasting AI progress. It is also likely that participation through all rounds, including through the discussion, which failed in this study, could be improved greatly through strong compensation of experts' time or through experts' participation as part of the administering organization.

8.9 Summary

This study made numerous novel contributions to the existing body of knowledge, all of which are detailed in the following chapter. The salient lessons are summarized in Table 16. These contributions all involved improving existing methods or the demonstration of newly proposed methods. Despite the breadth of techniques evaluated in this study there was still significant depth and rigor in the analysis and discussion of the results that were presented. The results discussed in this chapter can act as a guide for those interested in any aspect of forecasting AI progress or different levels of TAI. Even those who purport the utility of different methods to be superior to those here may learn from the notes in this chapter regarding evaluation of such a diverse set of methods applied to these ends. Forecasting progress in technology is hard enough that forecasters working to forecast a technology as complex and unique as an artificial form of intelligence itself should not discount any perspective due to the severity of the challenge[155].

[155] Private communications with one of the world's leading forecasters confirm as much.

Table 16: Salient Lessons from the Study

Component	no.	Most Significant Lessons Learned
Survey	1	Future of work researchers should not ignore possible extreme labor displacement from AI.
	2	Methods beyond surveys should be developed to forecast extreme labor displacement.
	3	Multivariate statistical models are useful for data analysis of quantile survey data.
	4	Further work should be conducted in pursuit of overconfidence controls for surveys.
Tech Mining	1	Tech mining is useful for identifying trajectories of emerging research topics.
	2	Stronger signal of emerging research trajectories can be obtained with richer, full-text data.
Extrapolation	1	Can be used to create logical forecasts for the development of DTAI and/or RTAI.
	2	Extrapolations are well suited for expert adjustment and a simple path to strong forecasts.
Structured Interviews	1	Participants take significantly more time to answer forecasting questions than for surveys.
	2	Interview participants exhibited significantly less overconfidence than survey participants.
	3	Can yield very useful qualitative data, particularly foresight over 18-24 month horizons.
Scenario Network Mapping	1	SNM is a very powerful technique, well-suited for mapping the paths to transformative AI.
	2	The success of this technique is very sensitive to the participants, workspace and scheduling.
	3	SNM is a very useful supplement for other scenario mapping techniques, e.g., JDM.
	4	Diversity of AGI experts is important, e.g. inclusion of bio-inspired and agentive AGI experts.
	5	Mapping the paths to AGI requires rigid facilitation by a skilled, knowledgeable facilitator.
Judgmental Distillation Mapping	1	JDM is a promising technique that is capable of quantitative forecasts of TAI/DTAI/RTAI.
	2	The Delphi may be a more suitable alternative for iterative elicitation than interviews.
	3	Aggregation of expert opinion during iterative expert elicitation is particularly challenging.
The Delphi	1	The Delphi technique is very promising for future variants of the holistic framework.
	2	The adapted Delphi process is well-suited for eliciting experts on apt forecasting targets.
	3	The research agenda created provides a solid framework for interested researchers.
Holistic Forecasting Framework	1	The holistic forecasting framework is a strong theoretical framework for TAI forecasting.
	2	Another example(s) of a holistic forecasting framework suggests its value for AI forecasts.
	3	Combining SNM and JDM through the holistic forecasting framework is very promising.
	4	Future work should focus on improved variants of the holistic forecasting framework.
	5	Future work can explore the incorporation of strategic planning elements to the framework.

9 CONTRIBUTIONS

This study contributes to the existing body of knowledge in numerous ways that can be classified into two primary categories: practical contributions and research contributions. The practical contributions of this research lie in the fact that no similar work has been conducted in the AI strategy domain, and that the work conducted in this study can be used to guide future efforts in developing practically useful forecasts to be used for planning purposes among policy-makers and governance researchers. The research contributions include the novel techniques incorporated to the numerous methods used here that can be harnessed by future researchers using the methods for any variety of applications. These classes of contributions are detailed in the lists below.

Practical contributions:

1. This study demonstrates the first survey to forecast extreme labor displacement from AI. It suggests that future of work models should consider extreme labor displacement scenarios from AI that are not considered in existing models.
2. This is the first study to use a rigorous approach, beyond simple roadmapping methods, to map the paths to AGI. The insights regarding the SNM workshopping process will be useful for organizations who may need to develop such maps.
3. Collectively, the contributions of the various novel methods and the insight into other techniques in the context of transformative AI are valuable to government agencies and think tanks that need to create these forecasts for informing policy initiatives.

Research contributions:

1. This study proposes scenario mapping, a new class of methods rooted in scenario planning methods that are necessary for forecasting and planning issues arising from complex, wicked problems. This can be useful for policy makers and future researchers to assist in distinguishing among different types of relevant methods for certain applications.
2. This study develops a new workshopping technique for mapping the AI technical landscape by using a scenario mapping technique (i.e., SNM), addressing the need identified by Dafoe (2018).
3. This study proposes and develops a new scenario mapping technique, judgmental distillation mapping, which offers another alternative solution for addressing the need identified by Dafoe (2018).
4. This study creates a new framework for forecasting TAI by combining scenario mapping techniques with numerous other techniques to form a holistic forecasting framework, addressing a need identified by Brynjolfsson and Mitchell (2017).
5. This study proposes another variation on the newly proposed holistic forecasting framework that is derived from a forecasting tournament framework proposed by Tetlock et al. (2017). This also addresses the need for a new framework described by Brynjolfsson and Mitchell (2017).
6. This study proposes an extensive framework for defining and understanding transformative AI and other related terms such as dramatically transformative AI and radically transformative AI which independently have substantive value due to their potential to be used by practitioners and AI policy professionals when discussing issues related to AI

policy. This addresses a need identified by gaps in the existing literature (Ayres 1990a, Ayres 1990b, Lipsey et al. 2005).

7. This study demonstrates a novel technique for eliciting expert opinion through a survey by using a naïve calibration. This addresses a gap in the existing literature.
8. This study demonstrates a multivariate statistical analysis of forecast probability distributions collected from survey data that improves upon previous models that have been used for AI forecasting. This addresses a gap in the existing literature.
9. This is the first study to explore the quality of AI forecasts using a mixed methods approach, also addressing a gap in the existing literature.
10. This is the first study to use the Delphi technique in the context of AI forecasting by using it to create a research agenda for AI forecasting, addressing a gap in the current literature, and to demonstrate its viability as a component of an alternative holistic forecasting framework.
11. The study produced a research agenda for future efforts to forecasting AI progress.

10 LIMITATIONS

This study was extremely broad and, while it made numerous significant contributions to the existing body of literature, it also had many shortcomings that are addressed briefly in this chapter. Because there are numerous significant components to this study, the chapter is broken down so that the limitations can be addressed for each one individually. The final section in this chapter addresses limitations of the study overall.

10.1 Practitioner Survey

Surveys for forecasting technological progress are difficult to execute, and, as the literature review revealed, the majority of those conducted for other technological forecasting topics involved small groups of experts. One limitation of this study was that we did not have access to enough experts for statistically significant samples of HLAI research experts and traditional AI or machine learning experts. Thus, we considered practitioners rather than experts. While this can be seen as a limitation, the expertise level of participants in previous AI "expert" surveys is questionable (Baum et al. 2012, Muller and Bostrom 2016), particularly when it comes to forecasting expertise. Moreover, no calibration questions or forecasting training was conducted to mitigate for overconfidence, thus the identification of such studies as "expert" surveys may give them more credibility than they deserve. This is certainly underscored by the discrepancy between human tasks forecasts and jobs forecasts found by Grace et al. (2018). Thus, the fact that practitioners were used in this survey rather than experts could be perceived to be a limitation as well as a strength.

There are numerous limitations to the practitioner survey conducted for this study, however, and one of the most significant is related to the notion of experts. Sample bias is a major limitation in this study because previous work has found there to be significant differences of opinion based on disciplines (Nordhaus 1994, Few et al. 2018), and researchers should attempt to represent all members of the research community when creating objective forecasts[156] (Morgan 2014). The practitioners who participated in the survey reported here had expertise or knowledge in the domain of artificial intelligence, and in many cases, their knowledge is limited to a specific subdomain of artificial intelligence. However, this was critical to the study, so that the impact of HLAI expertise could be evaluated as a factor impacting forecasts, thus it can be seen as a limitation regarding the quality of the forecasts, but a necessary one regarding the research question of the overall study. Sample bias was not limited to other disciplines altogether, but also involved the neglect of subdisciplines of AI research such as robotics and NLP. Roboticists' opinions would be particularly valuable when considering the automatability of human tasks, and no effort was made to ensure that their perspective was represented independently or proportionally to their composition in the populations being approximated.

Another major challenge in the development of the survey instrument (see Appendix A and Appendix B) used in this portion of the study involved the definitions of key terms. This is a challenge for any study relying on expert elicitation, and, while best efforts were made to define all terms very carefully, the definitions were still imperfect, and several elements were

[156] However, there is reason to question such research in the context of forecasting different levels of TAI. It has even been suggested that, for anticipating RTAI, only those who are working directly on the most advanced projects will be able to accurately anticipate these projects (Yudkowsky 2018). Anecdotal evidence of this comes from Enrico Fermi's anticipation of and participation in the development of nuclear fission: in 1939 Enrico Fermi told a colleague that the nuclear chain reaction was a remote possibility but only four years later Fermi was leading the development of the world's first nuclear reactor at the University of Chicago (Ord 2020). It is also interesting to consider that the world's foremost atomic scientist, Ernest Rutherford, declared the harnessing of atomic energy to be "moonshine" in 1933.

overlooked[157]. Particularly challenging terms to define were human tasks, and several comments complained about the definition used for this (see Table C2.1 in Appendix C concerning comments about ambiguity). While a significant effort was made to carefully define human tasks, no effort was made to define "feasibility" or "automatable" in the context of human tasks. This was an oversight, and a limitation. Another poorly defined term was "broadly capable system." The mixed methods study was conducted in part to explore the impacts of details of survey design and administration on the results of the survey, and, in this way can be thought of as a successful effort to counter the limitations posed by these definitions and the lack thereof. These limitations are discussed further in regard to the results and limitations of the mixed methods study.

Because the survey was designed to explore esoteric elements of the development of RTAI and thus different notions of AGI, there was one particularly unique limitation. In the final question, regarding the impact of unlimited compute on future AI research, the term unlimited was not defined and could be interpreted as infinite compute. For the practical purpose of determining whether AI practitioners perceived compute to be a bottleneck to AI research, this differentiation was not significant, but in the eyes of some particularly adept HLAI experts, unlimited compute understood as infinite compute would dramatically change the response to the question. This is the case considering asymptotically optimal algorithms and the AIXI agent which has been shown to be a theoretically optimal agent given infinite compute (Hutter 2005).

Perhaps the most significant limitation of the study concerned the scenarios of extreme labor displacement that were considered. Only five scenarios were considered because of attempts to avoid survey fatigue and to reduce the amount of time necessary to complete the survey. Furthermore, two of these forecasts were redundant in order to explore the effect of the type of

[157] This was the case despite testing the survey questions with a small group of experts.

system on the forecasts (i.e., narrow or broadly capable) and another was redundant given that the results for it were made irrelevant by the fixed years forecasting questions (i.e., the 10-year question). Forecasts for some of the 90% and 99% extreme labor-displacing AI scenarios were relatively near, e.g., 10 to 15 years, and it would be interesting to have included scenarios for 70% or 80% labor-displacing AI scenarios. This is a limitation and something that likely should be considered in future work.

There are a variety of other limitations. One was the lack of a complete calibration exercise (Morgan 2014). The naive calibration technique employed was an attempt to address this that seemingly fell short. The study also lacked a training module or instructions for how to improve forecasts and how to reduce overconfidence or other cognitive biases common among novice forecasters (Nemet et al. 2017). Yet another limitation lies in the order of the questions which, unintentionally, could have had the effect of priming participants for responses to the later questions.

10.2 Tech Mining

Because the objective of the tech mining demonstration was not ambitious the limitations were minimal. These limitations include the size of the datasets, the amount of time dedicated to analysis and the relatively standard techniques used. While the study showed the quality of the datasets to significantly impact the quality of the results, it did evaluate the impact of dataset size on the quality of the results. This is because the size of the datasets used, relative to the number of topics and the number of salient dimensions in the dataset, was small. Moreover, it is well known that dataset size for NLP scales much further than the size of the dataset included here (Kaplan et al. 2020). Thus, larger datasets should be used in future work when resources are available.

Clustering and topic modeling is a time consuming process, and better quality results could perhaps have been obtained given more time. However, the quality of the results was not a priority as the focus of this demonstration was to evaluate the suitability of the technique and the impact of the quality of data on the results. At the time this research was conducted, no aggregated data for tech mining was available publicly, so a significant amount of time was required to simply scrape and clean this data for analysis. Certainly, this is a timely process as the think tank tech mining project employing ~10 data scientists demonstrates.

Similarly, the techniques used were relatively standard and an extensive search of their tuning parameters and hyperparameters was not conducted due to time constraints. While t-SNE is a powerful technique, as are LDA and k-means, more advanced techniques using deep learning are also possible. Future work should explore some recent advances in unsupervised learning techniques for this purpose.

10.3 Indicator Extrapolation

Indicator extrapolation is such a straightforward forecasting technique that it is sometimes referred to as "naive forecasting" (Roper et al. 2011). The demonstration included as part of this study effectively applied the technique for generating a forecast for RTAI through analogous comparison with previous grand scientific endeavors. While this was effective, there were some limitations, including substantial uncertainty regarding the timeline and what in fact is being predicted.

An excellent example of indicator extrapolation in the context of AI forecasting is the example of Russell and Norvig (1995), who were able to use the ELO scores of chess playing algorithms to successfully predict that in 1997 an algorithm would be able to surpass the world champion's ELO score. Because Russell and Norvig identified what is considered a technological indicator, they were able to predict specific technological progress that the indicator was representative of. Because this demonstration used a social indicator, or more specifically, an economic indicator the predicted value was for levels of investment. Unlike the ELO score of AI chess systems, levels of investment do not translate to technological progress or performance. Rather, they can only be tied to levels of investment seen previously in grand scientific endeavors.

While the indicator that was identified is strong and can be thought of as an indicator of approaching grand scientific projects, it is still a limitation because it was a social indicator. Social indicators are always less desirable than technological indicators (Roper et al. 2011), however, they are often the best that can be found. If it was easy to identify a strong technological indicator of different levels of TAI, then this dissertation would be irrelevant, but it is not easy. Social indicators are also of great value not because they are able to indicate technological progress directly, but because extrapolation is such a strong forecasting technique. Even if they do not point directly at some future technological event or technology, they often do accurately forecast something of interest or relevance to the technology of interest. Here, they indicate that levels of investment equivalent to incredible historic scientific projects are something that private organizations may be approaching in the next half decade. While this does not specifically suggest that DTAI or RTAI will be created at a specific date, it can be seen to suggest that, if the trend continues – i.e., if the organization (Google) continues to see progress toward some objective that allows them to justify continuing to increase levels of investment into DeepMind – then we are likely heading toward DTAI or even RTAI.

Thus, despite not being able to forecast a technology or specific technological benchmark, DeepMind’s operating costs are a good indicator of progress toward something like DTAI or RTAI which are themselves not clearly defined technologies. Moreover, because DTAI and RTAI are closely tied to economic productivity and anomalies in metrics used to measure human progress, such as economic measures, the indicator identified here may be a better indicator for forecasting TAI/DTAI/RTAI than any technological indicator could be (it is unlikely that a technological indicator could be identified to forecast a specific technology with such profound implications for the economy). Because this is an extrapolated forecast, this forecast can be continually revised in light of updated data, so, despite the limitations of social indicators, this is still a very robust forecast for TAI/DTAI/RTAI. Furthermore, while it is thought of as a forecast for RTAI, it could also be thought of as an upper bound for a forecast of DTAI.

10.4 Mixed Methods Analysis

The mixed methods analysis conducted here is very insightful and valuable because it is the first study of its type for AI, and possible for technology forecasting more broadly. However, it does have many limitations. One significant limitation which was unable to be controlled for was the interview environment. The locations of interviews varied greatly, from over lunch, over coffee, standing in hallways of the conference center to subway transit and walking through Stockholm. This likely had an impact on some interviews, however, it was not perceived to have had such a significant effect so as to discard the data collected. Further limitations are also listed below, but those included here only represent the most significant limitations and are not comprehensive.

10.4.1 Quantitative Comparison with Survey

One of the most obvious limitations was the small size of the sample of experts who participated in the interviews. Although small sample size expert elicitations have been successful in the literature (Baker et al. 2009), in this context and in this mixed methods approach, the results were less than ideal. The small sample size precluded a comparison of the actual forecasts elicited from experts during interviews and those from the practitioners in the survey. This is perhaps the greatest shortcoming, and it would still be very valuable for a future study to compare the forecasts from a suitable sized sample of expert interviews with the results of an expert or practitioner survey also of a suitable size.

Another limitation in the existing quantitative analysis was the inability to better measure the amount of time practitioners spent on the survey as opposed to working on other tasks after having started the survey. Other platforms could make it possible to require the survey to be completed in a single sitting, but this could bias respondents to rush. There is no clear way to better assess the amount of critical thought expended in making the forecasts, but this would be a desirable area for improvement in future work if the intent was to assess the impact of elicitation technique (e.g. in-person interactive vs online).

10.4.2 Qualitative Elements

The lack of standard environment made assessing qualitative elements of the interviews more challenging. Some conference venues made finding quiet and desirable interview location easier, but no perfect interview environments were found. The best options were to visit onsite coffee shops during the most popular sessions. This was common for podcast producers, as was clear from the use of higher fidelity recording equipment during interviews that were able to be scheduled and located in such environments. The latest NeurIPS conference provided interview rooms for interviewers, however, the NeurIPS communications chair did not respond to repeated inquiries despite being very helpful the previous year.

Other limitations regarding the qualitative elements included the perceived need of the interviewer to rush through the initial qualitative questions in order to ensure that the quantitative questions could be completed in the 20-30 minute projected length of the interview. Including qualitative questions at the end of the interview would have alleviated this need and may not have negatively biased participants. Nearly all participants were content with longer discussions following the listed questions that exceeded the 30-minute estimate, but such cooperation was not assumed in order to be respectful of all participants' time. 20-30 minutes was selected because interviews over 30 minutes may have been seen as less desirable, and some participants were only able to allocate 20 minutes (despite all continuing past the stated 20 minute limit), so this was perceived to be effective at increasing participation rates despite the limitation it placed on the quality of responses to open-ended, qualitative questions.

10.5 Scenario Network Mapping

The developmental SNM workshops had many severe limitations that have already been highlighted in different Chapters of this study. Foremost of these were issues that were beyond the control of the study, such as the make of the various workshops' participants and the spaces available for conducting the workshop in. Efforts were made to make do with the resources and willing participants available, and the results are perceived to be satisfactory. Because the time necessary to conduct the workshop was so long (i.e., multiple days, and long sessions each day), it was hard to find suitable locations with pseudo experts willing to spend so much time developing a workshopping technique.

Despite the perception of having overcome such challenges, their severity certainly had an impact on the results. If given extensive resources to host experts and compensate them for their time to develop a new workshopping technique, or, if invited by an AGI research laboratory to develop the workshopping technique with their experts' participation and using their facilities, the results may have been much better, from a practical standpoint, at the cost of their experts' time (a significant amount of which would have been required for learning how best to manage and facilitate the workshops). From the perspective of learning about the method and developing a method that can be used to generate practically useful maps with such groups, the composition of the participants was not a limitation. However, the facilities used for the workshop were certainly a limitation and had a particularly negative effect on the results of the 2nd workshop. Had the lack of an adequate space been known prior to arrival in Marburg, an effort could have been made to rent an appropriate space. This would have been within the resource constraints for the project, but still may not have made a significant difference due to the motivation and poorly suited expertise of the participants.

While all three workshops were held internationally, all participants were fluent in English. However, the participants of the Moscow workshop had a preference for working in Russian, and at some points in the process Russian was used to facilitate more detailed discussions regarding technical aspects of the topic of discussion. This posed a challenge for facilitation because of the language barrier. The language barrier was also an issue in using NeyroKod, as all of the instructions in the application are in Russian. Thus, the language barrier was a significant limitation for the digitization of the results from the first two workshops due to a lack of expertise and usable instructions for the software used[158].

Another major limitation involving the first SNM workshop was the failure to require the participants to complete a survey. The participants were all asked to report their notes, but the response level for this request was very poor. Had participants completed the surveys, as the participants in the following two SNM workshops, a more thorough qualitative analysis could have been conducted regarding the similar and dissimilar experiences of the participants in the different workshops. This likely could have helped to shed more light on the reasons for why the first workshop was more successful.

[158] Google Translate was used to translate the instructions in the application, however, it seems much was lost in the translation.

10.6 Judgmental Distillation Mapping

JDM is a novel technique, and this study made a major contribution to the existing body of knowledge by introducing it and demonstrating its viability. However, there were still numerous limitations in the demonstration of this technique, as well as in the design of the technique itself. Much was learned in this exercise regarding the design, and some of these limitations are described further below. Limitations in the demonstration and evaluation of this novel method are also outlined.

Regarding the technique itself, much should be altered if the technique were to actually be implemented by an organization. For example, interviews are a limitation, and questionnaires would be a much more effective way to conduct the iterative expert elicitation component of the method. While this was not possible in this study because of the limited availability of experts, if the technique is actually to be used for generating quantitative forecasts then resources or other means should be required to ensure the availability of experts. Another limitation would be the lack of a team of expert forecasters with domain expertise. Again, if the technique were to actually be used to generate a forecast that was intended for decision makers and to be actionable, such a team would be necessary. This would be difficult to do because there are few experts in the world with domain expertise on the topic as well as expertise in forecasting, but generating a quantitative forecast on topics of TAI/DTAI/RTAI is a non-trivial task which involves extensive resources in order to be actionable.

Another limitation concerns the demonstration and evaluation employed here. Ideally, it would have been desirable to create some quantitative forecasts using the technique rather only demonstrating the viability as a qualitative means of generating a map. It would also be useful to evaluate which methods are best suited for use with the technique. The Delphi, prediction markets and forecasting tournaments are suggested, but even if an organization have the resources to implement the method rigorously enough to generate actionable forecasts, there is no information regarding which techniques would be best suited for determining the distributions for forecasts of the decomposed external nodes (much less for determining which such techniques would provide the greatest accuracy).

10.7 The Holistic Forecasting Framework

The holistic forecasting framework proposed here is another novel element of this study that is a significant contribution to the existing body of knowledge. However, it also has some shortcomings. Many of these came to light when conducting a proof of concept for this study. More broadly, the general notion of a holistic forecasting framework for forecasting TAI, or technologies generally, is relatively sound with no apparent limitations.

One of these limitations is how the technique, as originally described, overemphasized the need for an input mapping and input scenarios. This is likely not necessary at all, although other scenario mapping techniques can effectively augment the technique such as to improve the quality of the overall results. Moreover, it is likely that for most cases the use of anything other than another scenario mapping technique for generating an input is not helpful and is unnecessary information which can be burdensome.

The most glaring limitation of the holistic forecasting framework, is, like JDM, that it requires incredible resources to implement at a level of rigor that would be necessary for using it to generate forecasts for decision makers that were actionable and able to substantially reduce uncertainty. Again, like JDM, the only organization with the resources necessary to conduct the proposed holistic framework correctly would be governments, organizations or research labs, and very well-funded non-profit entities. Grants are not often able to provide the resources necessary for carrying out the proposed framework, and it certainly would not be possible outside of an academic institution with substantial enough clout to encourage the high levels of engagement from outside experts necessary for completing the process.

10.8 The Delphi

The Delphi was a very strong and well executed component of the study, however, there were still some oversights in the design and execution of the custom Delphi process for the specific application it was used for here. Perhaps most obviously, the failure to require answers for scores for each dimension for each of the questions and each of the methods which led to the use of multiple imputation for addressing issues caused by missing values. The questionnaire itself may have been poorly structured, as some participants questioned the value of the first question, which may have led to non-responses from other invitees. Furthermore, while significant effort was made to identify as many experts as possible who met the criteria that was used, some experts were still overlooked, particularly several experts working on issues related to the future of work.

In general, because the modified Delphi process used here was a novel adaptation of a policy Delphi, and because the technique itself was under development, some aspects of the design were difficult to execute. This was because most digital platforms are not designed to elicit information in the manner that was desirable for this study. This was the case for the attempt to facilitate discussion among participants regarding the results reported during the first round of the Delphi. This was also the case for the use of the spreadsheet for collecting the scoring information from participants. The instructions for the scoring were unclear to some participants, as three participants reached out via email to confirm whether 1 or 5 was associated with the highest value. This question was able to be answered, and careful inspection of the results given knowledge of each participants' background and preferences confirmed that all interpretations of the scale were consistent. However, other elements of the study for which the instructions were insufficient suffered.

The glaring example of flawed instructions was the failure to require an answer to all of the scores in the spreadsheet provided. This was a deliberate decision, intended to reduce survey fatigue effects so that the response rate would be maximal. Moreover, participants were explicitly instructed that it was not necessary to score each item. This appears to have had the effect of encouraging several participants to skip a large number of items, which made averaging the results challenging for topics with few responses. Had there been a manageable number of skipped items, which was what had been naively assumed, then descriptive statistics may have been sufficient without means to account for the impact of the missing data. However, this was not the case. In future work scores for each item could be required, or scores could be scaled from -5 to 5, allowing for a default of neutral on each question so that missing values could be avoided. If the latter was employed, instructions would have to make clear that the failure to answer was in fact a neutral response. This would likely work to avoid multiple imputation while also minimizing survey fatigue and non-responses.

Because several participants skipped a large number of responses, and because the sample size of respondents was relatively small, imputation was not an ideal option. Moreover, it was likely less effective than would be desirable for ranking/scoring topics in an expert research agenda, given the sensitivity and importance of these topics being reported publicly. However, this was the best that could be done given the situation, yet, due to this limitation the results can not be thought to be very robust.

The first question of the questionnaire was intended to determine whether participants felt AI forecasting to be a tractable problem. This was asked because AI forecasting is, for numerous reasons, more challenging than other technological forecasting which is already more challenging than most other types of economic forecasting. Some feel that the topic is particularly challenging, and that the best that can be done is to reduce uncertainty, and in cases of deep uncertainty, this involves simply removing portions of the tails of distributional forecasts. However, others believe strongly that forecasting is very tractable, and that AI forecasting, while posing unique challenges, is still just as tractable as any other type of forecasting. Consequently, some invitees may have found this question off-putting as the first question in the questionnaire. The question did seem appropriate as an icebreaker, but perhaps a question about participants' area of expertise or anything less controversial, would have been better for starting the questionnaire. The question would have been more palatable if placed later in the questions, and if it had perhaps been worded differently.

The final limitation of consequence involves the experts who were invited to participate in the study. One group was not included that would have met the criteria described in Chapter 4 regarding relevant expertise: those conducting future of work research. One economist working on this work was invited, but he did not respond, and two other participants have been involved in one related study, but their primary research relevant to the topic would not be considered future of work research. No attempts were made to reach out to others who have done significant work on this topic, and their opinions could have been very valuable. There are at least four experts who would qualify as experts for participation in the Delphi portion of this study who were not contacted. This is certainly a limitation, and, while another Delphi specifically for this purpose is unlikely anytime soon, future work should make extra efforts identify all researchers that might be relevant.

While there were several limitations, these limitations were beneficial for gaining experience using a modified version of the policy Delphi for identifying salient questions from highly regarded experts. As discussed earlier, this experience will be very valuable for the development of the technique in new contexts and applications.

10.9 Forecasting Transformative AI

The majority of the limitations have been covered in each of the previous subsections of this chapter. However, there were some limitations to this entire inquiry. The major limitation was that this inquiry attempted to be broad and comprehensive, so much of the project developed over the course of the initial portions of the study and thus was not entirely planned from the start. This is not an advisable course of action for most research questions, but for some of the most challenging research questions more unorthodox techniques are necessary. The research question of focus in this study of "How can we best forecast transformative AI?" is a terribly broad research question, and because of this it was difficult to develop a comprehensive plan for answering the question from the beginning. However, while such a broad research question poses significant challenges, it is important that such questions are not ignored, and consequently, the piecemeal manner in which this project was launched can be justified, particularly because the results offer substantial evidence and conclusions regarding the research question.

For others who attempt to shed light on such challenging questions in the future, this study can offer some guidelines. Unorthodox plans for resolving research questions are less than optimal and should be avoided if possible. If it is not possible to stick to tried and true methods of research, then the investigator must be truly motivated about the question of interest, confident in their ability to complete the inquiry and confident that a sufficient resolution exists. They also should have confidence in their decision making abilities and their forecasting abilities so that they are able to make such strategic planning decisions correctly. In short, selecting such broad and ambitious research questions is a risky endeavor and this should always be the first consideration when moving forward with research for which no clear path forward emerges at the beginning[159].

[159] For this study the first elements were the survey and tech mining. The survey was obviously not going to be a sufficient answer for the research questions, but, had more effort been spent on tech mining this study could have turned out very differently. While the fact that tech mining was not explored more may in some aspects be a limitation, it can also be thought of as a strength or virtue because spending further time on the one most widely used technique would have precluded the broad exploration of less commonly used judgmental forecasting techniques and novel scenario mapping techniques.

11 FORECASTS & FORESIGHT

Forecasters are well aware that they should be very careful when making public forecasts of future scenarios (Tetlock and Gardner 2016). However, with great caution we will attempt to summarize the results from the numerous components of this study in terms of their implications on what to expect from AI progress in the next decade. The details will be vague intentionally, but there are two key areas to focus on where experts seem to converge on substantial progress being made in the near future: robotics and natural language processing. There are also the two major forecasts from independent components of the study that have been previously reported that will again be highlighted here. These are addressed first.

11.1 Forecasts

There are two forecasts which were generated during this study that can be thought of as forecasts with significant limitations. The limitations of these forecasts have been discussed extensively in Chapters 8 and 10. Here, we briefly revisit these forecasts and their implications simply to be thorough considering that forecasting was the central concern of this study.

The first of these forecasts involves several different forecasts for different levels of extreme labor-displacing AI scenarios. These scenarios are representative of both DTAI and RTAI, and the survey results suggest that such scenarios could arrive between 10- and 50-years time (median forecasts at 50% probability). While there are very high levels of uncertainty attached to these forecasts, they are useful for helping to ascribe a non-trivial probability to dramatic or radical societal transformation driven by AI technologies. Particularly, these results seem to suggest that the 10-year and 15-year likely scenarios should not be ignored. A complete scenario, told with a relatively significant amount of detail, is narrated in Section 12.9.1 to give a better idea of how the development of such technologies could have a rapid and dramatic impact on society.

The second of these forecasts is a single indicator extrapolation. This also applies to both DTAI and RTAI, and suggests that within roughly four to six years some very powerful AI capabilities may be realized if the current levels of investment into such advanced AI systems is sustained. While this is only one indicator, it strongly suggests that something unanticipated by most researchers is attainable within the timeframe of a half-decade. The scenario described in Section 12.9.1 is another good example of how this could occur.

11.2 Foresight

During the course of this study much time was spent attending AI conferences, attending presentations on leading AI research efforts and speaking with researchers working on many different cutting-edge AI research projects. In the course of these interactions, two primary topics stood out as things to keep an eye on for potentially dramatic impacts on society and the economy in the next five years. These research areas – natural language processing and robotics – are discussed briefly below.

Progress in the past two years in the research area of NLP has been nothing short of remarkable. By many measures (Wang et al. 2018, Wang et al. 2019) it can be thought to be much more significant than progress in image processing between 2012 and 2014. There is also reason to believe that the progress will not slow anytime soon (Raffel et al. 2019, Brown et al. 2020, Lewis et al. 2020). Moreover, the progress in this area, unlike previous progress in image processing, has applications to a large number of business tasks that involve interaction with natural human language. One example could be thought to be the automation of call center operations, which does not seem too far off given the recent success and generalizability of advanced chatbots powered by such technologies (Adiwardana et al. 2020).

The other area where we might expect tremendous and dramatically transformative progress is robotics. Robotics is a very old field, but recent progress applying deep learning to some of the most challenging tasks is robotics, such as fine manipulation, has proven to be very successful. Moreover, a lot of research is being made on advanced theoretical methods that has yet to be operationalized and developed into a functional product. This may be because robotics development is more likely to follow traditional technology development paths, such as the levels of technological readiness.

12 CONCLUSIONS

The previous discussion and limitation chapters have covered a significant amount of the upshot from this study. However, in this chapter an attempt to summarize these conclusions in a broader context is made. We also attempt to summarize the highlights from each of the independent studies succinctly.

12.1 Practitioner Survey

While the survey was effective for generating forecasts for DTAI and RTAI, it is likely that this forecast is not good for anything more than reducing uncertainty in the tails of the distribution. Moreover, it is the conclusion of this study that surveys administered to a large number of experts are ineffective for generating actionable forecasts. However, the results of this survey were effective for identifying things for exploring in further detail in future work, and there are some applications for which surveys are suitable.

One application for which surveys are suitable may be for generating forecasts intended for public consumption. Due to risks of information hazards (Bostrom 2011), forecasts for different levels of TAI should be distributed cautiously, particularly when there is reason to believe that they are actionable. Like sensitive intelligence, such forecasts are likely to be considered classified by governments and would thus be closely guarded state secrets.

Surveys are also very effective for generating forecasts with limited resources, although limitations, such as the level of expertise of participants, are likely to be challenges. However, for applications that concern forecasting only TAI, and not DTAI or RTAI, surveys may be appropriate for organizations working on specific near-term transformative issues such as those related to topics such as facial recognition technology or lethal autonomous weapons. If used by such groups, it would be recommended to consider using structured interviews or to consider the limitations highlighted here[160].

12.2 Tech Mining

Based on the results of the tech mining demonstration in this study, it can be concluded that tech mining is a very promising technique for forecasting TAI and perhaps DTAI. The ability of tech mining to be applied for forecasts beyond the two to three-year range is unknown and is something that would be a good topic for future research. However, for near term forecasts, and for policy related issues, tech mining is one of the foremost tools in many technology forecasters' toolboxes and is a technique that should always be considered.

Furthermore, the results of this study lead to the conclusion that the higher the quality of the data the better the forecasts generated using tech mining techniques will be. This was demonstrated using similar sized datasets, but including one dataset that involved full text data whereas the other two datasets involved only abstracts. The implications of this for future applications of text mining in applications related to forecasting TAI are that significant efforts should be made to create the largest possible dataset including as much detail as possible. Such a dataset would include an incredibly wide variety of data from disparate data sources such as social media, data on authors, key figures, stakeholders, patent data, academic publications, etc. The generation of such a dataset would involve substantial resources and, outside of governments or large corporations, would be prohibitively expensive for all but the most well-funded non-profits (the resources are likely beyond what is available to academics).

[160] There would also be literature more relevant to surveys regarding policy issues that should likely be considered, as some of the more near-term TAI issues are entangled with other social issues and are not as strongly dependent on the technological progress as technologies associated with DTAI or RTAI.

12.3 Indicator Extrapolation

The resulting forecast from the demonstration of indicator extrapolation is more actionable[161] than the forecast generated using the survey. In general, the results from the demonstration of this technique in this study underscore how powerful the technique is. The data reported here depicts the most recent data that was reported for the 2018 operating costs during 2019, however, the plot had been generated without the most recent datapoint and had suggested a similar timeline when the data was extrapolated. Because the most recent datapoint confirmed the earlier forecast, this gives us more reason to treat it as actionable. If the datapoint that is made available this year continues to confirm the trajectory of DeepMind operating costs, then the forecast would become even more actionable to interested parties.

While actionable, this forecast does not paint a complete picture, and necessitates asking of further questions. Moreover, what it actually forecasts is the amount of investment being made in RTAI development, which, if it were to continue, would be a proxy for the confidence of ongoing research efforts to these ends. If this forecast were treated as actionable, it would be important to understand these nuances. Furthermore, this forecast should not be treated as actionable without consensus from experts after judgmental adjustment. These forecasts are not included in this study as this study is not intended to contain any forecasts which could be perceived to be purported as actionable.

12.4 Mixed Methods Analysis

The mixed methods portion of the study did not generate a practical forecast, nor was it successful at generating forecasts for comparison. However, while this would have been a desirable outcome, the intention was for exploring the differences between in-person interactive elicitation and online

[161] While more actionable, this forecast is not in fact actionable. This is discussed further at the end of the following paragraph.

elicitations. Furthermore, a secondary objective was to better understand nuances of elicitations specifically pertaining to AI. While not as successful as was possible, as discussed in Chapter 10, overall the implementation of this technique and this unique study design was effective at shedding light on some of the more opaque elements of expert elicitations for forecasting AI progress. Some particular results that should be considered are listed below[162]:

- Experts answering survey questions in interviews spent significantly more time than AI practitioners who completed survey questions online (based on assumptions about distractions during online competition: 0.024 >= p-value >= 0.000).
- While not conclusive, the results of this study appear to suggest that expert elicitations performed in-person as opposed to online induce more carefully thought through responses[163].
    - It is the conclusion of this study that in-person elicitations should be used whenever possible when the elicited information is critical, at least until further work suggests this to be unnecessary.
- Collectively, experts appear to have very prescient foresight over an 18 month to 24 month forecasting horizon.
- Semantic ambiguity is less of a problem in interviews than surveys, as is the terminology. However, because AI is a very unique topic, the terminology and semantics remain crucial in each interview because of the high potential value, especially when speaking with experts of unequal expertise in specific domains (e.g. their expertise can cause their perception of progress to diverge from an objective assessment).

[162] A more extensive list of salient lessons was included in Table 16.
[163] Significant potential for future work on this topic is suggested in Chapter 10.

12.5 Scenario Network Mapping

Both of the scenario mapping techniques that were tested are well-suited for applications related to forecasting TAI. SNM is qualitatively focused, and consequently it is better suited for more granular analysis of the different plausible paths and futures involving DTAI and RTAI. The three developmental workshops conducted to demonstrate the technique for this study were very useful for determining what is and what is not effective for the use of SNM in the context of AI, and more generally, for the use of SNM in the context of identifying plausible paths in the development of other technologies (as suggested by List (2006)).

The traditional SNM technique, utilizing tactile methods and a large wall space, is effective for the purpose of mapping the paths to RTAI (e.g., different notions of AGI). Moreover, the modified SNM workshopping process developed here can be successful with respect to a simple mapping of the technical paths to RTAI in fewer than four half-day workshops. However, to develop multiple layers that include scenarios representing the variety of coordination challenges that the different technological paths could bring to fruition would likely take longer than four half-day workshops, especially if the technical paths were also attempted to be mapped in a single process.

Based on the observations outlined in the previous paragraph, it is suggested that the development of a technical scenario network map be conducted separately from the development of different layers of nodes for this map. This has several benefits, including allowing for the use of different groups with different types of expertise for developing the different layers of the map as well as allowing for the use of a less rigid structure for each workshop that does not have to keep to a fixed time and which can be adjusted for the different elements of the map that are the focus of the current stage of development. Furthermore, this allows for iterative refinement of the map using different groups to build on previous work. The use of different groups with different types of expertise is also likely a positive feature for appealing to high-level decision makers (e.g., when weeks have been spent with different groups of experts that are part of an organization or that were sought out in a very intensive forecasting effort, decision makers are likely to give the results the maximum credibility).

These suggestions for the future application of the SNM technique in the context of forecasting different levels of TAI is in stark contrast to much that was suggested by List (2005) in his original description of the technique. However, Gaziulusoy demonstrated a very successful modification of the process (Gaziulusoy 2010), although not quite as dramatic as this modification, so such a modification can be expected to be successful here as well. Despite substantial work that remains to further refine the technique, the modification demonstrated here was found to be one of the most effective forms for forecasting different levels of TAI for the purpose of aiding decision makers, and it should be explored further and employed for this purpose in the future.

Another critical element of the use of SNM in this context involved the software used. The use of some type of software for digitizing the maps is necessary, and no widely available software options were found to be sufficient for this process. Thus, it is necessary to either use a software like that used here (i.e., NeyroKod), or to create the visualizations for the map manually with a software package that enables more artistic freedom and which is less structured. Ideally, the software package used here will continue to be improved and made more widely available with better support so that the process can remain as straightforward as possible because manual scenario network map creation can be challenging for a large number of nodes in a map.

The use of a specialized scenario mapping software, like NeyroKod, is also desirable for use in map generation. The development of an SNM workshopping process that relies entirely on this software will require significant effort to develop, however, this could pay large dividends. Some benefits include the ability to collaborate to generate a map remotely, which would enable collaboration of high-level experts at great distances, or the anticipated ability of the utilization of such a software to speed up the workshopping process. There are some challenges that could pose as a barrier to use in this manner, such as the fact that each expert would need to be proficient in the use of the software package being employed. However, this is something that could be overcome relatively easily, because the software package is not as difficult to use as some geometric design software packages, and the most important features of the software could likely be picked up in perhaps a few hours of tutorials if participants had some previous experience with other design software.

12.6 Judgmental Distillation Mapping

In contrast to SNM, JDM is a quantitatively focused forecasting technique, and consequently it is best suited for high level qualitative analysis that can be used as targets for a Bayes network of different future scenarios for DTAI or RTAI. However, like SNM, JDM is a very useful scenario mapping technique that is anticipated to be used in the context of forecasting different levels of TAI in the future. It is likely more appealing than SNM to many organizations because it is quantitative as well as quantitative, however, this should be thought to make it superior to SNM in regard to its usefulness for aiding decision makers on topics related to TAI/DTAI/RTAI. It is certainly wise to always be mindful of the limitations of the forecasts for the internal nodes in a judgmental distillation map.

While the conclusion of this study strongly supports the value of JDM and the practicality of its use in future efforts to forecast the various levels of TAI, it is important to note that the form of JDM proposed in Chapter 5 is limited. Only after addressing the limitations of JDM discussed in Chapter 10 should the technique be applied for practical applications. This includes the need for validating the quantitative elements of the forecasts for internal nodes in a judgmental distillation map using some toy problem, a non-trivial task in itself. Moreover, while the suggestions for improvements to the JDM process mentioned in Chapters 8 and 10 should all be heeded, they alone are insufficient for making the technique operational for generating forecasts of the quality necessary for informing critical decisions.

In general, both of the scenario mapping techniques proposed here have the potential to be very useful. They also have the potential to be complementary to each other, because of the focus of JDM on qualitative elements and the ability of JDM to be used for generating high-level estimates of timelines for broad technological capabilities. It is likely that they are both necessary as part of a holistic forecasting framework to provide the optimal information to future decision makers challenged with the tasks of planning for the future impacts of advanced AI.

12.7 The Delphi

The Delphi technique was successfully used to generate an objective list of both the most important research questions for forecasting AI progress as well as an objective list for the most important methods for prioritizing. Highlights from these results are depicted in Tables XII and XIII. Hopefully, these results, when published as a research agenda (Gruetzemacher et al. forthcoming), will act as a unifying document for this significant area of research. If successful, this could have substantial positive impact on future efforts for forecasting and strategic planning concerning various levels of TAI.

The Delphi was also successful in demonstrating a technique that is useful in the context of AI for identifying and prioritizing questions of interest to researchers. This is an incredibly valuable process given the fact that the Delphi study found that identifying the most important and significant forecasting targets was a top topic of interest to experts in AI forecasting. The use of the Delphi in this manner could easily be used to create an alternative holistic forecasting framework to the one proposed here, or, it could be used to modify Tetlock's (2017) holistic forecasting framework to incorporate expert elicitation. Either of these developments would be welcome, given the value there seems to be in taking a holistic perspective on the challenge of forecasting TAI.

12.8 Holistic Forecasting Framework

The holistic forecasting framework proposed here was a novel contribution to the existing body of knowledge, that, together with the proposed JDM technique, may be very useful in future work on these issues. The holistic framework goes beyond the simple implementation of the JDM process to include an iterative refinement process that could be used on a continuing basis to keep improving and updating forecasts for informing decision makers. Moreover, this is the only framework which has been demonstrated to function in this manner as a proof of concept.

The foremost contribution of the holistic forecasting framework discussed in this study is not the specific framework proposed, but the broader notion that a holistic perspective should be taken when attempting to solve the incredibly challenging tasks required to adequately plan and prepare for technological driven futures with so many unknowns such as those that would be associated with DTAI or RTAI. It is important that, rather than considering only the holistic forecasting framework proposed here, forecasters maintain a holistic forecasting mindset when approaching problems as challenging as those posed by the safe and beneficial development of advanced AI technologies such as DTAI or RTAI.

12.9 The Bigger Picture

All of the techniques presented in this study contribute to different items identified by the experts who participated in the Delphi process for generation of a research agenda (they addressed several of the top items). For this reason, this study can be thought of as a first step in the right direction for making progress on a very challenging task: reducing uncertainty regarding critical planning decisions for the development of TAI/DTAI/RTAI. However, there remains a vast amount of work to be done, much of which is suggested in Chapters 6, 7, 8 and 10.

Thus, the bigger picture is that, despite the breadth of the current study, it still only covers a small piece of what is necessary to effectively reduce uncertainty regarding critical decisions pertaining to different levels of TAI. As suggested by the general notion of a holistic forecasting framework, we will likely need to use all of the tools at our disposal to sufficiently address such challenging problems as those posed planning for the development of DTAI/RTAI. This study was designed to effectively begin progress toward this in numerous ways, and perhaps the research agenda (Gruetzemacher et al. forthcoming) will be a catalyst for progress in this direction.

In general, it is important to not take for granted the safe development of TAI/DTAI/RTAI. Those who study the development of nuclear weapons have realized how close we have come to catastrophic consequences in the short time, historically speaking, since their first use. The most well-known example of this are the actions of Stanislav Petrov on the 26th of September in 1983 when he was in command of a Soviet early warning system for intercontinental ballistic missiles that began indicating multiple incoming missiles from North America. This was during a time of heightened tensions between the United States and the Soviet Union, yet Petrov chose not to notify his superior officers. Had he notified his superior officers he could have commenced a chain of events that would have led to the unthinkable occurring. Thus, despite the relatively low loss of life that nuclear weapons have brought during their existence, we should not be complacent about the risks that experts do seem to agree that very powerful AI systems pose.

Another example of risks that are well known but that are commonly ignored or downplayed due to hyperbolic discounting common in politics are the risks posed by novel viruses which could lead to global pandemics. The risks for such pandemics were widely known to experts in public health, epidemiology and global catastrophic risk. However, not all nations were prepared for the outbreak of the 2019 novel coronavirus. Comparing the impact of the virus on the world's two foremost economies and superpowers, it is obvious as to which made prudent decisions informed by accurate models. Furthermore, based on the logistics and organization involved in Wuhan during the first quarter of 2020, it seems likely that robust scenario planning techniques were utilized for developing strategic plans. These strategic plans were ultimately successful in stopping the spread of COVID-19 in the world's most populous nation, while the same measures, enacted too late, led to much more tragic consequences in a country over 20 times smaller. This example is ongoing, and the final costs of the lack of well-informed strategic plans is not yet known, but it will certainly be far greater than the toll would have been had the proper techniques been used for planning and for informing prudent decision makers.

This example does not simply illustrate the power of effective forecasting, scenario planning and strategic planning, but it also demonstrates how significant it is to prepare for scenarios or events which are thought to be in the tails of distributions of risk (i.e., low probability events). This is particularly urgent in the case of TAI because there is the element of "lock-in," thus, mistakes from poor planning regarding the development of transformative technologies are unlikely to recede over time, such as the impact of a global pandemic. Thus, strategic plans for these low probability events is even more important. This case is only more extreme for increasingly transformative technologies, such as DTAI or RTAI, that have the potential to dramatically or radically transform society in an irreversible manner. It is not unreasonable to think that life as we know it could change over the course of weeks or months as it did during March of 2020 in the western world.

12.9.1 A Plausible Scenario

Consider a scenario involving the public release of a digital agent that could accomplish 90% of the tasks of assistants, analysts, programmers, etc. Further imagine that this new product was licensed at $10,000 per year and that it was available both on the cloud or on an organization's existing hardware. It is May of 2028, and the product was rolled out at a major event, similar to large events held by Google or Apple, and is available for purchase immediately, with an online and interactive demonstration of thousands of tasks comprising nearly 90% of the tasks of white collar employees – the agent itself is even able to host a virtual sales conference where it can demonstrate its abilities to be applied for the specific tasks your organization may be interested in automating.

Businesses begin buying the new agent on the day it is released and deploying them for a handful of positions for the rest of the work week. The new agents are each able to accomplish the work of roughly three people when running on organizations' existing hardware, but when running on the cloud, in an optimized environment, they are able to accomplish the work of eight people. They are also able to work for 24 hours a day, seven days a week. The productivity increase is so dramatic that organizations increase their original orders by an order of magnitude the following week, and by another order of magnitude the week after. Soon, many of the employees who had previously worked in roles of programmer, analyst or assistant are finding that they have little work to do. There are no plans in place for this scenario, and employers are not obligated to pay taxes on these new agents nor are they required to keep retain human employees for positions which become automated in a rapid fashion.

Organizations begin laying off a small portion of employees, just 2%. However, 2% of the 21.8 million in the United States employed in office and administrative support positions is 436,000 jobs. Things do not end here, though, and the layoffs are more dramatic for small businesses who, needing to cut costs to stay competitive, are forced to replace as much as 90% of these roles within a month. Yet these numbers have not considered the gig economy; the new agent is able to complete 25% of the tasks of workers employed in the gig economy, which is likely to exceed the 2017 estimate of 55 million in the United States. Within the weeks following the unveiling of this new agent, 10% of gig workers have lost their contracts, and this trend will only continue moving forward.

The sale of this new agent is not limited to the United States, however, and organizations globally are forced to act quickly as well to maintain a competitive edge. Because of a one-month free trial in developing nations, there are no risks for trying the new agent in many markets. By the end of a month's time, over 200,000,000 agents have been licensed. By the end of the next month, the organization producing these agents has a new revenue stream of two trillion dollars annually, which is nearly all income. Furthermore, this organization has not implemented a windfall clause (O'Keefe et al. 2020). Thus, the gains from the value of this new technology go entirely to the venture capitalists who had funded the stealth startup with only ten billion dollars. When the company goes public after three months, the valuation after the initial day of trading is sixty trillion dollars, which is equivalent to a six-thousand-fold return on investment.

In the following months, numerous organizations who failed to adopt the new technology quick enough are forced to file for bankruptcy while many other organizations in the AI product space are losing all of their previous revenue streams. Markets become a perpetual rollercoaster as investors' uncertainty about the future of tech giants, or any of the world's leading firms, remains constantly influx based rapidly evolving analysts' predictions and frequent press releases regarding strategic decisions for the new economic reality. Some of these tech companies are able to deploy technologies that are 80% as capable as the new agent, for roughly the same cost, but they are still unable to compete.

Such a scenario may sound radical, but this is in fact what would be considered a dramatically transformative AI technology. While the likelihood of such a technology occurring in the next four years may seem remote now, the forecasts from the survey administered in this study suggest a probability of greater than 1% of this occurring based on responses from all participants. Responses from HLAI participants suggest the likelihood of this occurring to be nearly 5%. These probabilities place the likelihood of this occurring, based on a survey of AI practitioners, to be in the range of the likelihood of a global pandemic such as that of COVID-19. These are probabilities which we cannot ignore and which we should adequately plan and prepare for.

12.9.2 Concluding Remarks

This study has demonstrated a variety of techniques that are suggested for use in forecasting and planning for the development of TAI/DTAI/RTAI. In doing this it has introduced many novel concepts including the definitions of TAI/DTAI/RTAI, as well as several methods and techniques which have been shown to be suitable for forecasting and planning for such advanced AI. Despite all of the contributions of this study, this is only a small step in the right direction. Because of the gravity of the implications of such advanced technologies, no time can be wasted in planning and preparing for what many feel to be low likelihood scenarios, but scenarios that many agree would have dramatic or radical impacts on society. This study will hopefully serve to motivate research toward these ends and to underscore the need for much more work on these topics.

REFERENCES

Adams, S., et al. (2012). "Mapping the landscape of human-level artificial general intelligence." AI Magazine 33(1): 25-42.

Agarwal, R. (2019). “Machine Learning and Enculturation: Perspective of International Human Rights in China.” IOSR Journal of Engineering. Available at SSRN: https://ssrn.com/abstract=3391858

Amer, M., et al. (2013). "A review of scenario planning." Futures 46: 23-40.

Amodei, D., et al. (2016). “Concrete problems in AI safety.” arXiv preprint arXiv:1606.06565.

Amodei, D. and Hernandez, D. (2018). AI and Compute. OpenAI (blog), OpenAI.

Appio, F.P., et al. (2017). “The light and shade of knowledge recombination: Insights from a general-purpose technology.” Technological Forecasting and Social Change, 125: 154-165.

Armstrong, J. S. (2001a). Principles of forecasting: a handbook for researchers and practitioners, Springer.

Armstrong, J. S. (2001b). “Extrapolation for time-series and cross-sectional data.” In Principles of forecasting (pp 217-243). Springer.

Armstrong, S. and K. Sotala (2015). “How we’re predicting AI–or failing to.” In Beyond artificial intelligence (pp. 11-29). Springer.

Armstrong, S., et al. (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.

Arrow, K. J., et al. (2008). “The promise of prediction markets.” American Association for the Advancement of Science.

Aspinall, W. (2010). "A route to more tractable expert advice." Nature 463(7279): 294.

Autor, D. H., et al. (2003). "The skill content of recent technological change: An empirical exploration." The Quarterly journal of economics 118(4): 1279-1333.

Avin, S., et al. (2020). “Exploring AI Futures Through Role Play.” In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Axelrod, R. (2015). Structure of decision: The cognitive maps of political elites, Princeton university press.

Ayres, Robert U. (1990a). “Technological Transformations and Long Waves. Part I.” Technological Forecasting and Social Change 37(1): 1-37.

Ayres, Robert U. (1990b). “Technological Transformations and Long Waves. Part II.” Technological Forecasting and Social Change 36(1): 111-137.

Bacci, M.L., (2017). A concise history of world population. John Wiley & Sons.

Badia, A.P., et al. (2020). "Agent57: Outperforming the atari human benchmark." arXiv preprint arXiv:2003.13350.

Baker, E., et al. (2009). "Advanced solar R&D: Combining economic analysis with expert elicitations to inform climate policy." Energy Economics 31: S37-S49.

Baker, E., et al. (2010). "Battery technology for electric and hybrid vehicles: Expert views about prospects for advancement." Technological Forecasting and Social Change 77(7): 1139-1146.

Baker, E. and Keisler, J.M. (2011). "Cellulosic biofuels: Expert views on prospects for advancement." Energy 36(1): 595-605.

Bahdanau, D., et al. (2014). "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473.

Bauer, Z. (2018). AI Index Annual Report. AI Index.

Baum, S. D., et al. (2011). "How long until human-level AI? Results from an expert assessment." Technological Forecasting and Social Change 78(1): 185-195.

Beard, S. et al. (2020). "An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards." Futures 115.

Berry, B.J., et al. (1993). “Are long waves driven by techno-economic transformations?: Evidence for the US and the UK.” Technological forecasting and social change, 44(2): 111-135.

Bistline, J. (2014). "Energy technology expert elicitations: an application to natural gas turbine efficiencies." Technological Forecasting Social Change 86: 177-187.

Blei, D.M., et la. (2003). “Latent dirichlet allocation.” Journal of machine Learning research, 3(1): 993-1022.

Bocquet-Appel, J.P., 2011. “When the world’s population took off: the springboard of the Neolithic Demographic Transition.” Science 333(6042): 560-561.

Bostrom, N., 2011. “Information hazards: a typology of potential harms from knowledge.” Review of Contemporary Philosophy (10): 44-79.

Bostrom, N. (2014). Superintelligence, Dunod.

Bradfield, R., et al. (2005). "The origins and evolution of scenario techniques in long range business planning." Futures 37(8): 795-812.

Bresnahan, T.F., and Trajtenberg, M. (1995). “General Purpose Technologies ‘Engines of growth’?.” Journal of Econometrics 65(1): 83-108.

Brier, G.W. (1950). “Verification of forecasts expressed in terms of probability.” Monthly weather review 78(1): 1-3.

Brown, N. and Sandholm, T. (2017). “Libratus: The Superhuman AI for No-Limit Poker. International Joint Conference on Artificial Intelligence.

Brown, N. and Sandholm, T. (2019). Superhuman AI for multiplayer poker. Science, 365(6456): 885-890.

Brown, Tom B., et al. (2020). "Language models are few-shot learners." arXiv preprint arXiv:2005.14165.

Brundage, M. (2016). “Modeling progress in AI.” Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.

Brynjolfsson, E. and T. Mitchell (2017). "What can machine learning do? Workforce implications." Science 358(6370): 1530-1534.

Brynjolfsson, E., et al. (2018). “What can machines learn, and what does it mean for occupations and the economy?” AEA Papers and Proceedings.

Brynjolfsson, E., et al. (2019). Artificial Intelligence and the Modern Productivity Paradox. National Bureau of Economic Research.

Carey, R. (2018). Interpretting AI and Compute Trends. AI Impacts (blog), AI Impacts.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies, Cambridge University Press.

Chermack, T. J. (2011). Scenario planning in organizations: how to create, use, and assess scenarios, Berrett-Koehler Publishers.

Cireşan, D. C., et al. (2010). "Deep, big, simple neural nets for handwritten digit recognition." Neural computation 22(12): 3207-3220.

Ciresan, D. C., et al. (2011). “Flexible, high performance convolutional neural networks for image classification.” Twenty-Second International Joint Conference on Artificial Intelligence.

Clemen, R. T. and R. L. Winkler (1999). "Combining probability distributions from experts in risk analysis." Risk analysis 19(2): 187-203.

Colson, A. R. and Cooke, R. M. (2018). "Expert elicitation: using the classical model to validate experts’ judgments." Review of Environmental Economics and Policy 12(1): 113-132.

Cooke, R. (1991). Experts in uncertainty: opinion and subjective probability in science. Oxford University Press.

Creswell, J. W. (2014). Research design: Qualitative, quantitative, and mixed methods approaches, Sage publications.

Curtright, A. E., et al. (2008). Expert assessments of future photovoltaic technologies. ACS Publications.

Czaplicka-Kolarz, K., et al. (2009). "Technology foresight for a vision of energy sector development in Poland till 2030. Delphi survey as an element of technology foresighting." Technological Forecasting and Social Change 76(3): 327-338.

Dafoe, A. (2018). AI Governance: A Research Agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford.

Dahmen, R., et al. (2008). “Delphi process yielded consensus on terminology and research agenda for therapeutic footwear for neuropathic foot.” Journal of clinical epidemiology 61(8): 819-e1.

Das, S., et al. (2020). “Learning Occupational Task-Shares Dynamics for the Future of Work.” Proceedings of the AAAI/ACM Conference on AI for Ethics, and Society.

David, H. (2013). The "task approach" to labor markets: an overview. No. w18711. National Bureau of Economic Research.

David, P.A. (1985). “Clio and the Economics of QWERTY.” The American Economic Review 75(2): 332-337.

Davison, R., et al. (2004). "Principles of canonical action research." Information systems journal 14(1): 65-86.

De Fauw, J., et al. (2018). “Clinically applicable deep learning for diagnosis and referral in retinal disease.” Nature medicine, 24(9): 1342-1350.

Deng, J., et al. (2009). "Imagenet: A large-scale hierarchical image database." IEEE conference on computer vision and pattern recognition.

Dimmitt, C., et al. (2005). “Identifying a school counseling research agenda: A Delphi study.” Counselor Education and Supervision, 44(3), 214-228.

Ding, J. (2018). Deciphering China’s AI dream. Future of Humanity Institute Technical Report.

Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books.

Drexler, K. E. (2019). Reframing Superintelligence. Oxford University, Oxford University.

Drucker, P. (1965). "The first technological revolution and its lessons." Technology and Culture 7(2): 143-151.

Duckworth, P., et al. (2019). "Inferring Work Task Automatability from AI Expert Evidence." 2nd AAAI/ACM Conference on AI for Ethics and Society.

Duin, P. A. (2006). Qualitative futures research for innovation, Eburon Uitgeverij BV.

Ecoffet, A., et al. (2019). "Go-Explore: a New Approach for Hard-Exploration Problems." arXiv preprint arXiv:1901.10995.

Ecoffet, A., et al. (2020). "First return then explore." arXiv preprint arXiv:2004.12919.

Efron, B. and R. J. Tibshirani (1994). An introduction to the bootstrap. CRC press.

Etzoni, O. (2020). “How to know if artificial intelligence is about to destroy civilization.” MIT Technology Review.

Everitt, T., et al. (2018). “AGI safety literature review.” arXiv preprint arXiv:1805.01109.

Farmer, J. D. and F. Lafond (2016). "How predictable is technological progress?" Research Policy 45(3): 647-665.

Fernández-Macías, E., et al. (2018). "A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work." arXiv preprint arXiv:1807.02416.

Fisher, J.C. and Pry, R.H., (1971). “A simple substitution model of technological change.” Technological forecasting and social change, 3, pp.75-88.

Fowler Jr, F. J. (2013). Survey research methods, Sage publications.

Fountaine, T., (2019). “Building the AI-powered organization.” Harvard Business Review, pp.63-73.

Frank, M. R., et al. (2019). "Toward understanding the impact of artificial intelligence on labor." Proceedings of the National Academy of Sciences 116(14): 6531-6539.

Fraser-Mackenzie, P., et al. (2015). “The prospect of a perfect ending: Loss aversion and the round-number bias.” Organizational Behavior and Human Decision Processes, 131, pp.67-80.

Frey, C. B. and Osborne, M. A. (2017). "The future of employment: how susceptible are jobs to computerisation?" Technological Forecasting and Social Change 114: 254-280.

Few, S., et al. (2018). "Prospective improvements in cost and cycle life of off-grid lithium-ion battery packs: An analysis informed by expert elicitations." Energy Policy 114: 578-590.

Garcia, M. L. and Bray, O. H. (1997). Fundamentals of technology roadmapping. Sandia National Labs., Albuquerque, NM (United States).

Garfinkel, B. (2018). “Reinterpretting AI and Compute.” AI Impacts (blog), AI Impacts.

Gaziulusoy, A. (2010). “System innovation for sustainability: a scenario method and a workshop process for product development teams.” The University of Auckland.

Gaziulusoy, A., et al. (2013). "System innovation for sustainability: a systemic double-flow scenario method for companies." Journal of Cleaner Production 45: 104-116.

Gil, Y. and Selman, B. (2019). A 20-Year Community Roadmap for Artificial Intelligence Research in the US. AAAI and ACM report. arXiv preprint arXiv:1908.02624.

Gill, I. (2020). “Whoever leads in artificial intelligence in 2030 will rule the world until 2100.” The Brookings Institute, Washington DC.

Glenn, J. C. (2006). "Nanotechnology: Future military environmental health considerations." Technological Forecasting and Social Change 73(2): 128-137.

Goertzel, B., (2014a). “Artificial general intelligence: concept, state of the art, and future prospects.” Journal of Artificial General Intelligence, 5(1), pp.1-48.

Goertzel, B. (2014b). Ten Years to the Singularity If We Really, Really Try. Humanity+ Press.

Goertzel, B. (2016). The AGI Revolution: An Inside View of the Rise of Artificial General Intelligence. Humanity+ Press.

Goldstein, J.S., (2006.) “The Predictive Power of Long Wave Theory, 1989-2004” In Kondratieff Waves, Warfare and World Security (pp.137). IOS Press.

Goodier, C., et al. (2010). "Causal mapping and scenario building with multiple organisations." Futures 42(3): 219-229.

Gordon, S. C., & Barry, C. D. (2006). “Development of a school nursing research agenda in Florida: A Delphi study.” The Journal of school nursing 22(2): 114-119.

Gordon, T. J. and O. Helmer (1964). Report on a long-range forecasting study. RAND Corporation.

Grace, K. (2013). Algorithmic progress in six domains. Machine Intelligence Research Institute.

Grace, K. (2015). "AI Timeling Surveys." AI Impacts (blog), AI Impacts. https://aiimpacts.org/ai-timeline-surveys/.

Grace, K., et al. (2018). "When will AI exceed human performance? Evidence from AI experts." Journal of Artificial Intelligence Research (62): 729-754.

Graves, A. (2012). "Sequence transduction with recurrent neural networks." arXiv preprint arXiv:1211.3711.

Green, W. H. (2000). Econometric Analysis (5th ed.), New York: Prentice Hall.

Green, K.C., et al. (2015). Methods to elicit forecasts from groups: Delphi and prediction markets compared. Foresight.

Grinin, L.E., et al. (2017). "Forthcoming Kondratieff wave, Cybernetic Revolution, and global ageing." Technological Forecasting and Social Change, 115: 52-68.

Gruetzemacher, R. (2018). "Rethinking AI Strategy and Policy as Entangled Super Wicked Problems." 2018 AAAI/ACM Conference on AI for Ethics and Society. New Orleans, LA.

Gruetzemacher, R., et al. (2018). "3D deep learning for detecting pulmonary nodules in CT scans." Journal of the American Medical Informatics Association, 25(10), pp.1301-1310.

Gruetzemacher, R. (2019a) "A Holistic Framework for Forecasting Transformative AI." *Big Data and Cognitive Computing* 3(3): 35.

Gruetzemacher, R. (2019b). "2018 Trends in DeepMind Operating Costs (updated)." (blog) http://www.rossgritz.com/uncategorized/updated-deepmind-operating-costs/

Gruetzemacher, R. and Paradice, D. (2019a) "Alternative Techniques for Mapping Paths to HLAI." *arXiv preprint arXiv:1905.00614*.

Gruetzemacher, R. and Paradice, D. (2019b) "Toward Mapping the Paths to AGI." International Conference on Artificial General Intelligence. Springer.

Gruetzemacher, R. and Whittlestone, J. (2019). "Defining and Unpacking Transformative" AI. arXiv preprint arXiv:1912.00747.

Gruetzemacher, R., et al. (forthcoming). "Forecasting AI Progress: A Research Agenda."

Gruetzemacher, R., Paradice, D., Lee, K. B. (2020; forthcoming). “Forecasting Extreme Labor Displacement: A Survey of AI Practitioners.”

Hahn, R. W. and P. C. Tetlock (2005). "Using information markets to improve public decision making." Harvard Journal of Law and Public Policy 29: 213.

Harris, K., et al. (2018). "Labor 2030: The collision of demographics, automation and inequality." Bain & Company.

He, K., et al. (2016). “Deep residual learning for image recognition.” Proceedings of the IEEE conference on computer vision and pattern recognition.

Henningsen, A. and J. D. Hamann (2007). "systemfit: A package for estimating systems of simultaneous equations in R." Journal of Statistical Software 23(4): 1-40.

Hernández-Orallo, J. (2014). "AI Evaluation: past, present and future." arXiv preprint arXiv:1408.6908.

Hernández-Orallo, J. (2017). "Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement." Artificial Intelligence Review 48(3): 397-447.

Hinton, G. E., et al. (2006). "A fast learning algorithm for deep belief nets." Neural computation 18(7): 1527-1554.

Hochreiter, S. and Schmidhuber, J., (1997). “Long short-term memory.” Neural computation, 9(8):1735-1780.

Hora, S. C., et al. (2013). "Median aggregation of distribution functions." Decision Analysis 10(4): 279-291.

Hutter, M. (2004). Universal artificial intelligence: Sequential decisions based on algorithmic probability, Springer Science & Business Media.

Inayatullah, S. (1998). "Causal layered analysis: Poststructuralism as method." Futures 30(8): 815-829.

Jetter, A. and W. Schweinfort (2011). "Building scenarios with Fuzzy Cognitive Maps: An exploratory study of solar energy." Futures 43(1): 52-66.

Jetter, A. J. and K. Kok (2014). "Fuzzy Cognitive Maps for futures studies—A methodological assessment of concepts and methods." Futures 61: 45-57.

Jonnes, J. (2004). Empires of light: Edison, Tesla, Westinghouse, and the race to electrify the world. Random House Trade Paperbacks.

Kaplan, Jared, et al. (2020). "Scaling laws for neural language models." arXiv preprint arXiv:2001.08361.

Karnofsky, Holden. (2016). “Some Background on our Views Regarding Advanced Artificial Intelligence. Blog.” Open Philanthropy Project (blog), Open Philanthropy Project. https://www.openphilanthropy.org/blog/some-background-our-views-regarding-advanced-artificial-intelligence

Kassie, M., et al. (2013). "Adoption of interrelated sustainable agricultural practices in smallholder systems: Evidence from rural Tanzania." Technological Forecasting and Social Change 80(3): 525-540.

Kellum, J. A., et al. (2008). “Development of a clinical research agenda for acute kidney injury using an international, interdisciplinary, three-step modified Delphi process.” Clinical Journal of the American Society of Nephrology, 3(3), 887-894.

Kondratiev, Nikolay. (1926). Die Langen Wellen der Konjunktur. (The long waves of the economy). Archiv fur Sozialwissenschaft und Sozialpolitik (56), 573.

Kosko, B. (1986). "Fuzzy cognitive maps." International journal of man-machine studies 24(1): 65-75.

Krizhevsky, A., et al. (2012). “Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems.

Kucharavy, D. and De Guio, R. (2008). Logistic substitution model and technological forecasting. TRIZ Future 2008 - Synthesis in Innovation. HAL Archives-Ouvertes.

Kvale, S. and S. Brinkmann (2009). Interviews: Learning the craft of qualitative research interviewing. Sage.

Lafond, F., et al. (2018). "How well do experience curves predict technological progress? A method for making distributional forecasts." Technological Forecasting and Social Change 128: 104-117.

Lake, B. M., et al. (2017). "Building machines that learn and think like people." Behavioral and brain sciences 40.

Le, Q.V. (2013). "Building high-level features using large scale unsupervised learning." 2013 IEEE international conference on acoustics, speech and signal processing.

LeCun, Y., et al. (1989). "Backpropagation applied to handwritten zip code recognition." Neural computation 1(4): 541-551.

LeCun, Y., et al. (1998). "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86(11): 2278-2324.

LeCun, Y., et al. (2015). "Deep learning." Nature 521(7553): 436.

Legg, S. and Hutter, M. (2007a). “A collection of definitions of intelligence.” Frontiers in Artificial Intelligence and applications, 157, p.17.

Legg, S. and Hutter, M. (2007b). "Universal intelligence: A definition of machine intelligence." Minds and machines 17(4): 391-444.

Leike, J., et al. (2018). “Scalable agent alignment via reward modeling: a research direction.” arXiv preprint arXiv:1811.07871.

Lewis, Patrick, et al. (2020). "Retrieval-augmented generation for knowledge-intensive nlp tasks." arXiv preprint arXiv:2005.11401.

Liebowitz, S.J. and Margolis, S.E. (1995). “Path Dependence, Lock-in, and History.” Journal of Law, Economics, & Organization. 205-226.

Linstone, H.A. and Turoff, M. (1975). The delphi method (pp. 3-12). Addison-Wesley.

Lipinski, A. and D. Loveridge (1982). "Institute for the future's study of the UK, 1978–1995." Futures 14(3): 205-239.

Lipsey, R. (2007). “Transformative Technologies in the Past Present and Future: Implications for the US Economy and US Economic Policy.” A presentation for the ITIF Breakfast Forum.

Lipsey, R., et al. (2005). Economic transformations: general purpose technologies and long-term economic growth. Oxford University Press.

List, D. (2005). Scenario network mapping: the development of a methodology for social inquiry. University of South Australia Adelaide, SA, Australia.

List, D. (2006) "Action research cycles for multiple futures perspectives." Futures 38.6: 673-684.

List, D. (2006) Scenario Network Mapping User Manual.

List, D. (2007). "Scenario network mapping." Journal of Futures Studies 11(4): 77-96.

Magee, C., et al. (2016). "Quantitative empirical trends in technical performance." Technological Forecasting and Social Change 104: 237-246.

Manyika, J., (2017). A future that works: AI, automation, employment, and productivity. McKinsey Global Institute Research, 60.

Manning, C.D., and Schütze, H. (1999). Foundations of statistical natural language processing. MIT press.

Marchau, V.A., et al. (2019). Decision making under deep uncertainty. Springer.

Marcus, G., and Davis, E. (2019) Rebooting AI: building artificial intelligence we can trust. Pantheon.

Martínez-Plumed, F. (2018). “Accounting for the neglected dimensions of ai progress.” arXiv preprint arXiv:1806.00610.

Martínez-Plumed, F. and Hernández-Orallo, J. (2018.) “Analysing Results from AI Benchmarks: Key Indicators and How to Obtain Them.” arXiv preprint arXiv:1811.08186.

Martínez Plumed, F., et al. (2020). “Does AI Qualify for the Job? A Bidirectional Model Mapping Labour and AI Intensities.” AAAI/ACM Conference on AI for Ethics and Society.

Maaten, L.V.D. and Hinton, G. (2008). “Visualizing data using t-SNE.” Journal of machine learning research, 9(11): 2579-2605.

Martino, J.P. (1993). Technological forecasting for decision making. McGraw-Hill.

McKay, C., et al. (2019). Automation and a Changing Economy. The Aspen Institute.

McKinney, S.M., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788): 89-94.

McWaters, J.R., (2018). The New Physics of Financial Services: Understanding how artificial intelligence is transforming the financial ecosystem. Deloitte.

Mensch, G., (1979). Stalemate in technology: innovations overcome the depression. Ballinger Publishing Company.

Mensch, G.O., et al. (1987). “Outline of a formal theory of long-term economic cycles.” In The Long-Wave Debate (pp. 373-389). Springer.

Michie, D. (1973). "Machines and the theory of intelligence." Nature 241(23.02): 1973.

Mikolov, T., et al. (2013). “Distributed representations of words and phrases and their compositionality”. Advances in neural information processing systems.

Minsky, M. L., et al. (2004). "The St. Thomas common sense symposium: designing architectures for human-level intelligence." AI Magazine 25(2): 113-113.

Mnih, V., et al. (2013). "Playing atari with deep reinforcement learning." arXiv preprint arXiv:1312.5602.

Møldrup, C., et al. (2001). "Risks of future drugs: A Danish expert Delphi." Technological Forecasting and Social Change 67(2-3): 273-289.

Moore, D.A., et al. (2017). “Confidence calibration in a multiyear geopolitical forecasting competition.” Management Science, 63(11): 3552-3565.

Moore, G. E. (1965). “Cramming more components onto integrated circuits.” Electronics 38 (8).

Morgan, M. G. (2014). "Use (and abuse) of expert elicitation in support of decision making for public policy." Proceedings of the National Academy of Sciences 111(20): 7176-7184.

Mosekilde, E., and Rasmussen., S. (1986). "Technical economic succession and the economic long wave." European Journal of Operational Research 25(1): 27-38.

Mozur, P. (2017). Beijing Wants AI to be Made in China by 2030. New York Times.

Muehlhauser, L. (2015a). Human progress before the industrial revolution. (blog) https://lukemuehlhauser.com/human-progress-before-the-industrial-revolution/

Muehlhauser, L. (2015b). What do we know about AI timelines? Open Philanthropy Project (blog), Open Philanthropy Project.

Muehlhauser, L. (2016). “What should we learn from past AI forecasts?” Open Philanthropy Project (blog), Open Philanthropy Project

Muehlhauser, L., and Sinick, J. (2014). “How Big is the Field of Artificial Intelligence? (initial findings).” Machine Intelligence Research Institute (blog), Machine Intelligence Research Institute.

Muehlhauser, L. (2017). “There was only one industrial revolution.” (blog) https://lukemuehlhauser.com/there-was-only-one-industrial-revolution/

Müller, V. C. and N. Bostrom (2016). “Future progress in artificial intelligence: A survey of expert opinion.” In Fundamental issues of artificial intelligence, (pp. 555-572) Springer.

Mullins, C. (2012). Retrospective analysis of technology forecasting: In-scope extension. The Tauri Group.

Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Nagy, B., et al. (2011). "Superexponential long-term trends in information technology." Technological Forecasting and Social Change 78(8): 1356-1364.

Nagy, B., et al. (2013). "Statistical basis for predicting technological progress." PloS one 8(2): e52669.

Nakicenovic, N., (1979). Software package for the logistic substitution model. International Institute for Applied Systems Analysis.

Nakicenovic, N., (1987). “Technological substitution and long waves in the USA.” In The long-wave debate (pp. 76-103). Springer, Berlin, Heidelberg.

Nemet, G. F., et al. (2017). "Quantifying the effects of expert selection and elicitation design on experts’ confidence in their judgments about future energy technologies." Risk Analysis 37(2): 315-330.

Nordhaus, W. D. (1994). "Expert opinion on climatic change." American Scientist 82(1): 45-51.

North, D. C., and Thomas, R. P. (1977). “The first economic revolution.” The Economic History Review, 30(2), 229-241.

Noyes, Jan. (1983). “The QWERTY Keyboard: A Review.” International Journal of Man-Machine Studies 18(3): 265-281.

Ng, A. (2018). AI Transformation Playbook: How to Lead Your Company Into the AI Era. Landing AI.

O'Keefe, Cullen, et al. (2020). "The Windfall Clause: Distributing the Benefits of AI for the Common Good." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.

Omar, M., et al. (2017). "Global mapping of artificial intelligence in Google and Google Scholar." Scientometrics 113(3): 1269-1305.

Ord, T. (2020). The Precipice. Hachette Books.

Papageorgiou, E. I. (2013). Fuzzy cognitive maps for applied sciences and engineering: from fundamentals to extensions and learning algorithms, Springer Science & Business Media.

Pennington, J., et al. (2014). “Glove: Global vectors for word representation.” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Perla, P.P. and McGrady, E.D. (2011). “Why wargaming works.” Naval War College Review, 64(3), pp.111-130.

Perrault, R., et al. (2019). The AI Index 2019 Annual Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Phaal, R., et al. (2004). "Technology roadmapping—a planning framework for evolution and revolution." Technological Forecasting and Social Change 71(1-2): 5-26.

Pichai, S., and Schwab, K. (2020). “DAVOS 2020 | An Insight, An Idea with Sundar Pichai.” World Economic Forum. Davos, Switzerland. https://www.youtube.com/watch?v=7sncuRJtWQI

Porter, A. L. and S. W. Cunningham (2004). Tech mining: exploiting new technologies for competitive advantage, John Wiley & Sons.

PwC. (2018). “The Macroeconomic Impact of Artificial Intelligence.” https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf.

Radford, A., et al. (2019). “Language Models are Unsupervised Multitask Learners.” OpenAI (blog), OpenAI.

Rea, L. M. and R. A. Parker (2014). Designing and Conducting Survey research: A Comprehensive Guide, John Wiley & Sons.

Roper, A. T., et al. (2011). Forecasting and Management of Technology, 2nd edition, John Wiley & Sons.

Rowe, G. and G. Wright (2001). “Expert opinions in forecasting: the role of the Delphi technique.” Principles of Forecasting, (pp. 125-144) Springer.

Roxburgh, C. (2009). "The use and abuse of scenarios." McKinsey Quarterly 1(10): 1-10.

Rumelhart, D. E., et al. (1985). Learning internal representations by error propagation. University of California at San Diego, La Jolla Institute for Cognitive Science.

Rumelhart, D. E., et al. (1988). "Learning representations by back-propagating errors." Cognitive Modeling 5(3): 1.

Russell, S. J. and P. Norvig (1995). Artificial intelligence: a modern approach. First edition.

Russell, S. J. and P. Norvig (2016). Artificial intelligence: a modern approach, Third edition. Pearson Education Limited.

Ruttan, V.W., (2006). Is war necessary for economic growth?: military procurement and technology development. Oxford University Press.

Sanders, N.R. and Ritzman, L.P., (2001). “Judgmental adjustment of statistical forecasts.” In Principles of Forecasting (pp. 405-416). Springer.

Schmidhuber, J. (2015). "Deep learning in neural networks: An overview." Neural Networks 61: 85-117.

Schoemaker, P. J. (1995). "Scenario planning: a tool for strategic thinking." Sloan Management Review 36(2): 25-41.

Schoemaker, P. J. and P. E. Tetlock (2016). "Superforecasting: How to upgrade your company's judgment." Harvard Business Review 94: 72-78.

Schumpeter, Joseph Alois. 1939. Business Cycles. Vol. 1. McGraw-Hill.

Searle, J. R. (1980). "Minds, brains, and programs." Behavioral and Brain Sciences 3(3): 417-424.

Sejnowski, T.J., 2020. "The unreasonable effectiveness of deep learning in artificial intelligence." Proceedings of the National Academy of Sciences.

Shapiro, Carl, and Hal R. Varian. (1998). Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press.

Shu, C. (2014). Google Acquires Artificial Intelligence Startup DeepMind for over $500 Million. TechCrunch. https://techcrunch.com/2014/01/26/google-deepmind/

Silver, D., et al. (2016). "Mastering the game of Go with deep neural networks and tree search." Nature 529(7587): 484.

Silver, D., et al. (2017a). "Mastering chess and shogi by self-play with a general reinforcement learning algorithm." arXiv preprint arXiv:1712.01815.

Silver, D., et al. (2017b). "Mastering the game of go without human knowledge." Nature 550(7676): 354.

Simon, H.A., 1996. Models of my life. MIT press.

Soares, N. and Fallenstein, B. (2017). "Agent foundations for aligning machine intelligence with human interests: a technical research agenda." In The Technological Singularity (pp. 103-125). Springer, Berlin, Heidelberg.

Socher, R., et al. (2013). "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the 2013 conference on empirical methods in natural language processing.

Soetanto, R., et al. (2011). "Unravelling the complexity of collective mental models: a method for developing and analysing scenarios in multi-organisational contexts." Futures 43(8): 890-907.

Sutton, R.S. and Barto, A.G. (2018). Reinforcement learning: An introduction. MIT press.

Sutton, Richard. (2019). The Bitter Lesson. Incomplete Ideas (blog). http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Szegedy, C., et al. (2015). "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition.

Taylor, J., et al. (2016). “Alignment for advanced machine learning systems.” Machine Intelligence Research Institute.

Tetlock, P. and D. Gardner (2016). Superforecasting: The Art and Science of Prediction, Random House.

Tetlock, P. E. (2006). Expert Political Judgment: How Good Is It? How Can We Know?-New Edition, Princeton University Press.

Tetlock, P.E., (2017). Full-inference-cycle tournaments. A grant proposal to the Intelligence Advanced Research Planning Activity (unfunded).

Tolman, E. C. (1948). "Cognitive maps in rats and men." Psychological Review 55(4): 189.

Tseng, F.-M., et al. (2009). "Assessing market penetration combining scenario analysis, Delphi, and the technological substitution model: the case of the OLED TV market." Technological Forecasting and Social Change 76(7): 897-909.

Turchin, Alexey. (2018). “Could Slaughterbots Wipe Out Humanity? Assessment of the Global Catastrophic Risk Posed by Autonomous Weapons.” URL: https://philpapers.org/rec/TURCSW

Turoff, M., (1970). “The design of a policy Delphi.” Technological Forecasting and Social Change, 2(2), pp.149-171.

Twiss, B.C., (1986). Managing technological innovation. Longman Publishing Group.

Van Duijn, J. J. (2013). The long wave in economic life. Routledge.

van Neuss, L. (2015). Why did the Industrial Revolution Start in Britain? Available at SSRN 2696076.

Varho, V., et al. (2016). "Futures of distributed small-scale renewable energy in Finland—A Delphi study of the opportunities and obstacles up to 2025." Technological Forecasting and Social Change 104: 30-37.

Vaswani, A., et al. (2017). “Attention is all you need.” In Advances in neural information processing systems (pp. 5998-6008).

Vinyals, O., et al. (2019). AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. DeepMind (blog), Google DeepMind.

Walsh, T. (2018). "Expert and non-expert opinion about technological unemployment." International Journal of Automation and Computing 15(5): 637-642.

West, Darrel M., Allen, John R. April 24, 2018. How Artificial Intelligence is Transforming the World. Report, The Brookings Institute, Washington, D. C.

White House., 2018, May. Summary of the 2018 White House Summit on Artificial Intelligence for American Industry. In United States. Office of Science and Technology Policy. United States. Office of Science and Technology Policy.

Wiser, R., et al. (2016). "Expert elicitation survey on future wind energy costs." Nature Energy. 1(10): 1-8.

Yudkowsky, E. (2017). “There’s No Fire Alarm for Artificial General Intelligence.” Available online: https://intelligence.org/2017/10/13/fire-alarm/ (accessed on 31 May 2019).

Yun, C., et al. (2019). “Are Transformers universal approximators of sequence-to-sequence functions?” arXiv preprint arXiv:1912.10077.

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." Journal of the American statistical Association 57(298): 348-368.

Zhang, B. and A. Dafoe (2019). "Artificial Intelligence: American Attitudes and Trends." Available at SSRN 3312874.

APPENDIX A: SURVEY QUESTIONS (SWEDEN)

Welcome (pg 1)

Welcome to our survey on forecasting AI progress. You have been selected as a participant of ICML, IJCAI or HLAI. We thank you for your participation! As a token of our appreciation you will receive a $10 virtual Visa card or Amazon gift card upon completion. All of your information will remain confidential. More information on the study and your privacy can be found in the official information letter.

This complete survey is estimated to take 8 minutes to complete. It consists of 13 questions on 6 pages and 1 page with 2 simple basic information questions. Click OK to begin.

Basic Information (pg 2)

1. What is your first name and last name?
   first name ____________________
   last name ____________________

2. What is your employer's name and your job title?
   employer's name ____________________
   job title ____________________

Calibration (pg 3)

INSTRUCTIONS:

These questions concern autonomous vehicles and the levels of autonomy for autonomous vehicles. Please consider the following:

- Level 4 vehicles are fully autonomous without driver intervention for a majority of driving conditions.

- Level 5 vehicles are fully autonomous without driver intervention for all driving conditions. Such vehicles would not require a steering wheel.

For each of the scenarios listed these questions please select the probability that you assign to it happening.

3. Level 4 autonomous vehicles are available to US consumers by the end of 2019.
________ (select a probability from 0% to 100%)

4. More than one corporation produces software used in Level 5 autonomous vehicles by the end of 2021.
_______ (select a probability from 0% to 100%)

Calibration (pg 4)

5. OpenAI has stated that they intend to beat professional players in the video game Dota 2 next month. What probability do you assign for that happening?
_______ (select a probability from 0% to 100%)

6. In 2018 DeepMind has already published one article in the journal Nature and one article in the journal Science. What probability do you assign to DeepMind publishing at least one more article in equal or better quality journals?
_______ (select a probability from 0% to 100%)

(pg 5)

INSTRUCTIONS:

These questions will ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

7. What percentage of human tasks do you believe that it is feasible for current AI technology to perform at or above the level of a typical human?
_______ (select a percentage from 0% to 100%)

8. In 5 years, what percentage of human tasks do you believe that it will be feasible for AI technology to perform at or above the level of a typical human?
_______ (select a percentage from 0% to 100%)

9. In 10 years, what percentage of human tasks do you believe that it will be feasible for AI technology to perform at or above the level of a typical human?

_______ (select a percentage from 0% to 100%)

10. Comments (optional):

_______ (large text box)

(pg 6)

INSTRUCTIONS:

These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

11. In how many years do you expect AI systems to collectively be able to accomplish 50% of human tasks at or above the level of a typical human? Think feasibility.

_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

12. Comments (optional):

_______ (large text box)

(pg 7)

INSTRUCTIONS:
These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will also ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

These questions will also refer to a broadly capable AI system. By this we mean a single implementation of an algorithm that can accomplish a large variety of tasks. For example, if a single implementation of an algorithm could complete 50% of human tasks we would consider it a broadly capable AI system.

13. In how many years do you expect AI systems to collectively be able to accomplish 90% of human tasks at or above the level of a typical human? Think feasibility.

_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

14. In how many years do you expect a broadly capable AI system to be able to accomplish 90% of human tasks at or above the level of a typical human?

_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

15. Comments (optional):

_______ (large text box)

(pg 8)

INSTRUCTIONS:

These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will also ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

These questions will also refer to a broadly capable AI system. By this we mean a single implementation of an algorithm that can accomplish a large variety of tasks. For example, if a single implementation of an algorithm could complete 50% of human tasks we would consider it a broadly capable AI system.

16. In how many years do you expect AI systems to collectively be able to accomplish 99% of human tasks at or above the level of a typical human? Think feasibility.

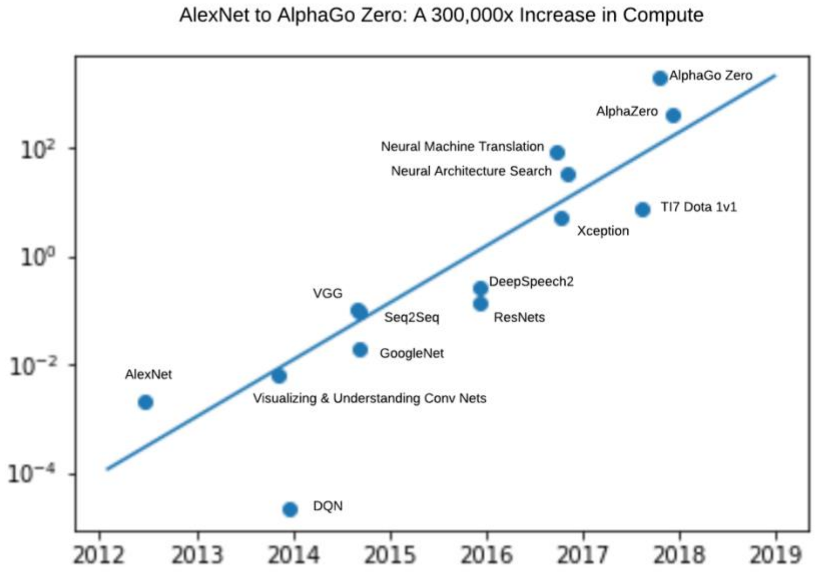
_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

17. In how many years do you expect a broadly capable AI system to be able to accomplish 99% of human tasks at or above the level of a typical human?
_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

18. Comments (optional):
_______ (large text box)

INSTRUCTIONS:

These questions all concern compute. The first two refer to the figure below, created this year by OpenAI. Please consider this figure when answering these next two questions.



(Figure A.1: AI and Compute)

19. What do you believe is the probability that the trend in the figure continues for major milestones and achievements for the next 5 years?
_______ (select a probability from 0% to 100%)

20. What do you believe is the probability that the trend in the figure continues for major milestones and achievements for the next 10 years?

_______ (select a probability from 0% to 100%)

21. How much do you believe progress in AI would accelerate if all researchers had access to unlimited compute? Please select the best answer.

_______ It would slow down
_______ 0%
_______ 25%
_______ 50%
_______ 100%
_______ 200%
_______ 400%

22. Comments (optional):

_______ (large text box)

APPENDIX B: SURVEY QUESTIONS (PRAGUE)

Welcome (pg 1)

Welcome to our survey on forecasting AI progress. You have been selected as a participant of ICML, IJCAI or HLAI. We thank you for your participation! As a token of our appreciation you will receive a $10 virtual Visa card or Amazon gift card upon completion. All of your information will remain confidential. More information on the study and your privacy can be found in the official information letter.

This complete survey is estimated to take 8 minutes to complete. It consists of 13 questions on 6 pages and 1 page with 2 simple basic information questions. Click OK to begin.

Basic Information (pg 2)

1. What is your first name and last name?
   first name ____________________
   last name ____________________

2. What is your employer's name and your job title?
   employer's name ____________________
   job title ____________________

Calibration (pg 3)

INSTRUCTIONS:

These questions concern autonomous vehicles and the levels of autonomy for autonomous vehicles. Please consider the following:

- Level 4 vehicles are fully autonomous without driver intervention for a majority of driving conditions.

- Level 5 vehicles are fully autonomous without driver intervention for all driving conditions. Such vehicles would not require a steering wheel.

For each of the scenarios listed these questions please select the probability that you assign to it happening.

3. Level 4 autonomous vehicles are available to US consumers by the end of 2019.
________ (select a probability from 0% to 100%)

4. More than one corporation produces software used in Level 5 autonomous vehicles by the end of 2021.
_______ (select a probability from 0% to 100%)

Calibration (pg 4)

5. In the 2016 US elections significant efforts were made by malign actors to influence the outcome. What probability do you assign to deep learning being used to generate fake video and audio for the purpose of influencing the 2018 US national elections on November 6th?
_______ (select a probability from 0% to 100%)

6. In 2018 DeepMind has already published one article in the journal Nature and one article in the journal Science. What probability do you assign to DeepMind publishing at least one more article in equal or better quality journals?
_______ (select a probability from 0% to 100%)

(pg 5)

INSTRUCTIONS:

These questions will ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

7. What percentage of human tasks do you believe that it is feasible for current AI technology to perform at or above the level of a typical human?
_______ (select a percentage from 0% to 100%)

8. In 5 years, what percentage of human tasks do you believe that it will be feasible for AI technology to perform at or above the level of a typical human?

_______ (select a percentage from 0% to 100%)

9. In 10 years, what percentage of human tasks do you believe that it will be feasible for AI technology to perform at or above the level of a typical human?

_______ (select a percentage from 0% to 100%)

10. Comments (optional):

_______ (large text box)

(pg 6)

INSTRUCTIONS:

These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

11. In how many years do you expect AI systems to collectively be able to accomplish 50% of human tasks at or above the level of a typical human? Think feasibility.

_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

12. Comments (optional):

_______ (large text box)

(pg 7)

INSTRUCTIONS:

These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will also ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

These questions will also refer to a broadly capable AI system. By this we mean a single implementation of an algorithm that can accomplish a large variety of tasks. For example, if a single implementation of an algorithm could complete 50% of human tasks we would consider it a broadly capable AI system.

13. In how many years do you expect AI systems to collectively be able to accomplish 90% of human tasks at or above the level of a typical human? Think feasibility.
_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

14. In how many years do you expect a broadly capable AI system to be able to accomplish 90% of human tasks at or above the level of a typical human?
_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

15. Comments (optional):
_______ (large text box)

(pg 8)

INSTRUCTIONS:

These questions will require you to make forecasts about future AI progress. Each question will ask for your forecasts at probabilities of 10%, 50% and 90%.

These questions will also ask your opinion of future AI progress with regard to human tasks. We define human tasks as tasks that humans are currently paid to do. We consider human tasks as different from jobs in that an algorithm may be able to replace humans at some portion of tasks a job requires while not being able to replace humans for all of the job requirements. For example, an algorithm(s) may not replace a lawyer entirely but may be able to accomplish 50% of the tasks a lawyer typically performs.

These questions will also refer to a broadly capable AI system. By this we mean a single implementation of an algorithm that can accomplish a large variety of tasks. For example, if a single implementation of an algorithm could complete 50% of human tasks we would consider it a broadly capable AI system.

16. In how many years do you expect AI systems to collectively be able to accomplish 99% of human tasks at or above the level of a typical human? Think feasibility.

_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

17. In how many years do you expect a broadly capable AI system to be able to accomplish 99% of human tasks at or above the level of a typical human?

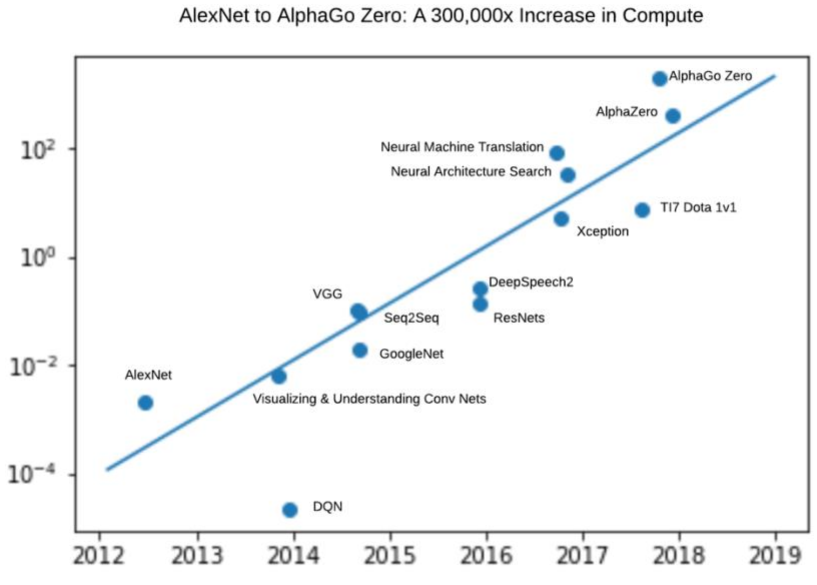
_______ 10% probability of this occurring
_______ 50% probability of this occurring
_______ 90% probability of this occurring

18. Comments (optional):

_______ (large text box)

INSTRUCTIONS:

These questions all concern compute. The first two refer to the figure below, created this year by OpenAI. Please consider this figure when answering these next two questions.



(**Figure B.1:** AI and Compute)

19. What do you believe is the probability that the trend in the figure continues for major milestones and achievements for the next 5 years?

_______ (select a probability from 0% to 100%)

20. What do you believe is the probability that the trend in the figure continues for major milestones and achievements for the next 10 years?

_______ (select a probability from 0% to 100%)

21. How much do you believe progress in AI would accelerate if all researchers had access to unlimited compute? Please select the best answer.

_______ It would slow down
_______ 0%
_______ 25%
_______ 50%
_______ 100%
_______ 200%
_______ 400%

22. Comments (optional):

_______ (large text box)

APPENDIX C: EXTRA SURVEY RESULTS

This section details the results from the extreme labor displacement survey conducted at the 2018 International Conference on Machine Learning (ICML), the 2018 International Joint Conference on Artificial Intelligence (IJCAI) and the 2018 Human-level Artificial Intelligence Conference (HLAI-conf). For further reading, see Gruetzemacher et al. 2020.

C1 Participants' Demographics and Comments

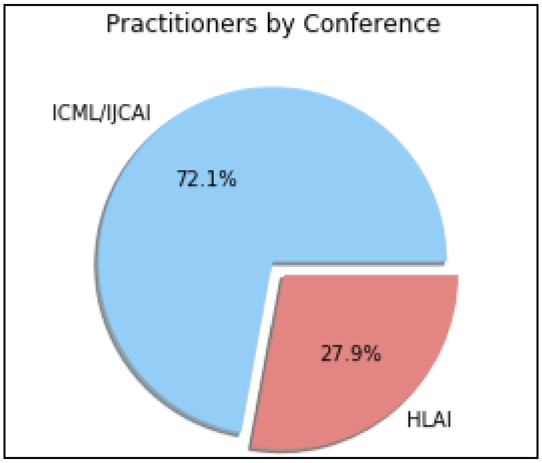


Figure C1.1: Participants by conference location.

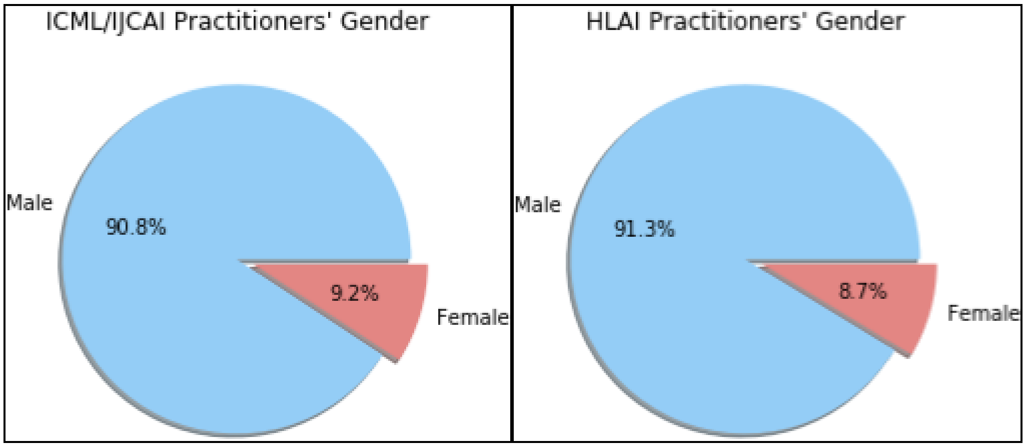


Figure C1.2: AI practitioner survey participation by gender.

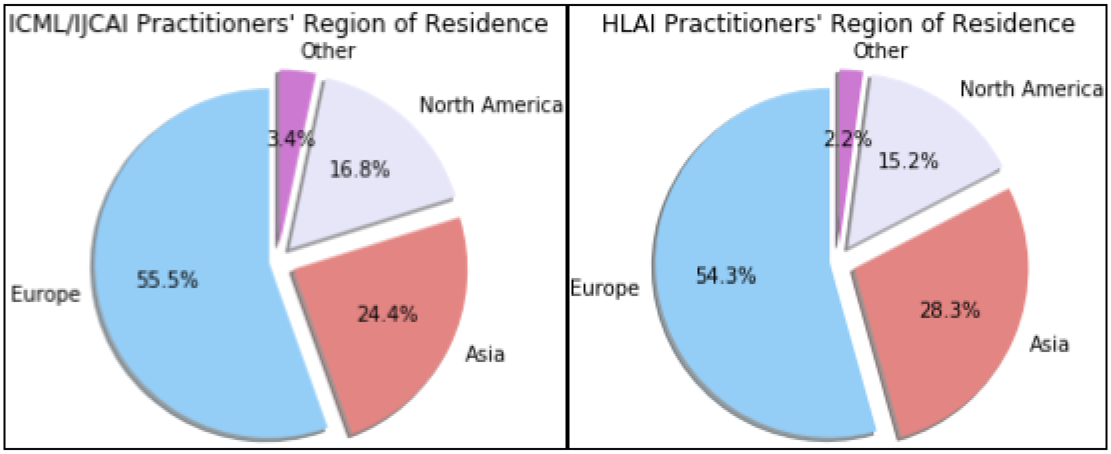


Figure C1.3: AI practitioner survey participation by region of residence.

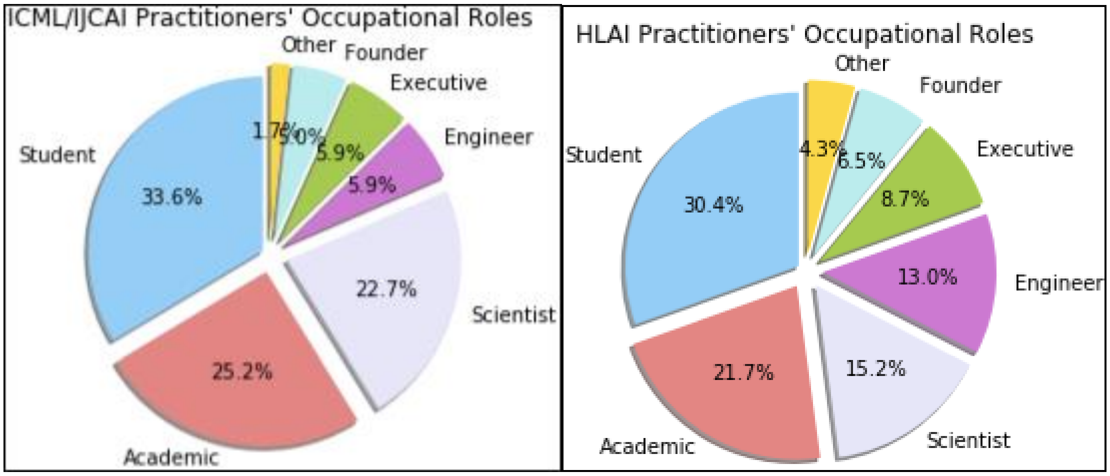


Figure C1.4: AI practitioner survey participation by occupational role.

Table C2.1: Analysis of Survey Respondents' Comments

	Question Topic				
Comment Type	Fixed Years	50%	90%	99%	Compute
Comments to consider further	6	8	9	8	5
Other comments	11	1	2	3	10
Confusion due to ambiguity	1	3	4	3	4
Difficulties in estimating	3	2	2	3	3
Never or very long time	n/a	1	3	5	n/a

C2 Conference Comparison Plots for Extreme Labor Displacement Scenarios

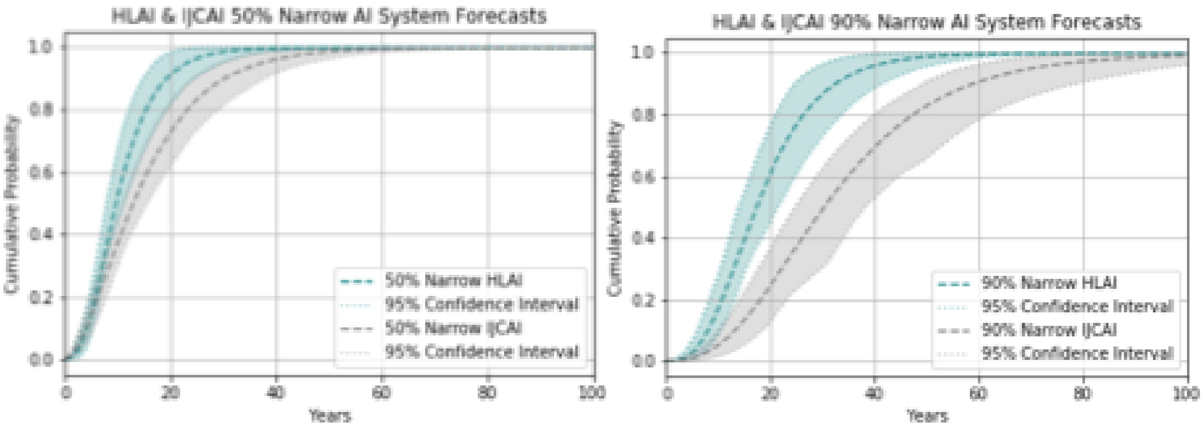


Figure C2.1: Comparisons of forecasts from participants of HLAI and participants of ICML/IJCAI (IJCAI); 50% narrow system (left) and 90% broadly capable system (right).

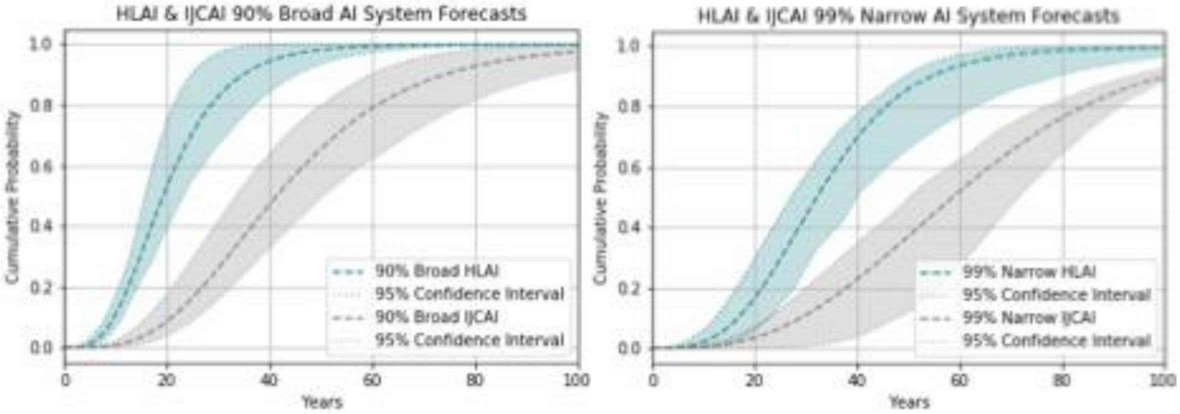


Figure C2.2: Comparisons of forecasts from participants of HLAI and participants of ICML/IJCAI (IJCAI); 90% broadly capable system (left) and 99% narrow system (right).

APPENDIX D: EXTRA TECH MINING RESULTS

This appendix contains the complete tech mining results visualizations. The two cases shown in Chapter 6 do not appear here. The remaining four are included in Figures D.1, D.2, D.3 and D.4.

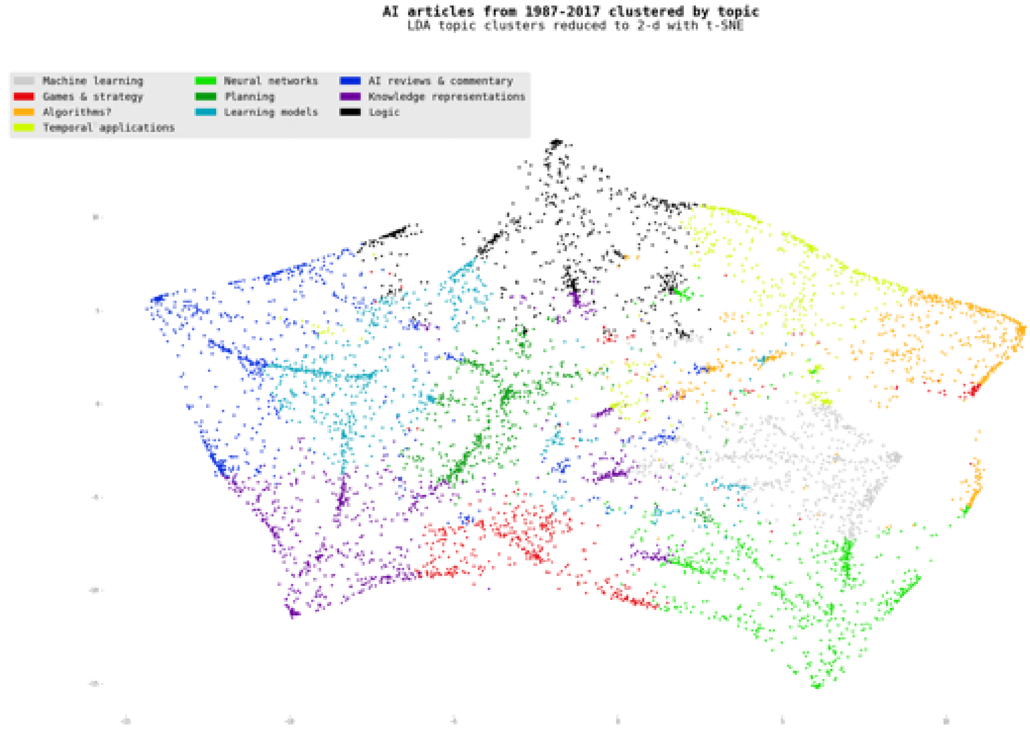


Figure D.1: LDA topics visualized with t-SNE for the Web of Science AI dataset.

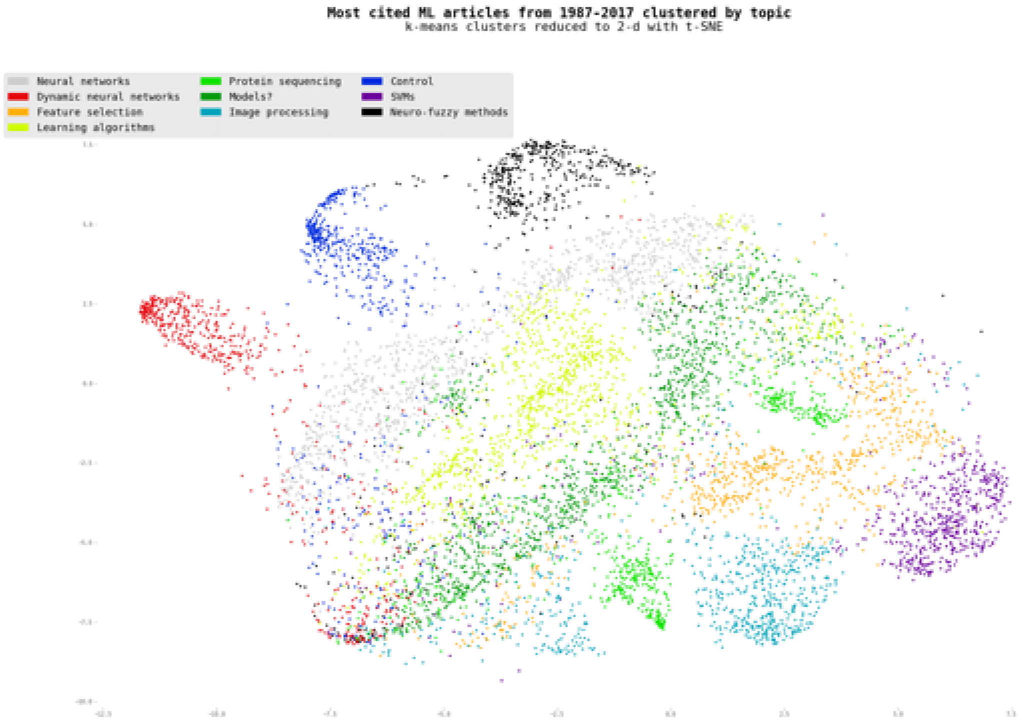


Figure D.2: k-means clusters visualized with t-SNE for the Web of Science ML dataset.

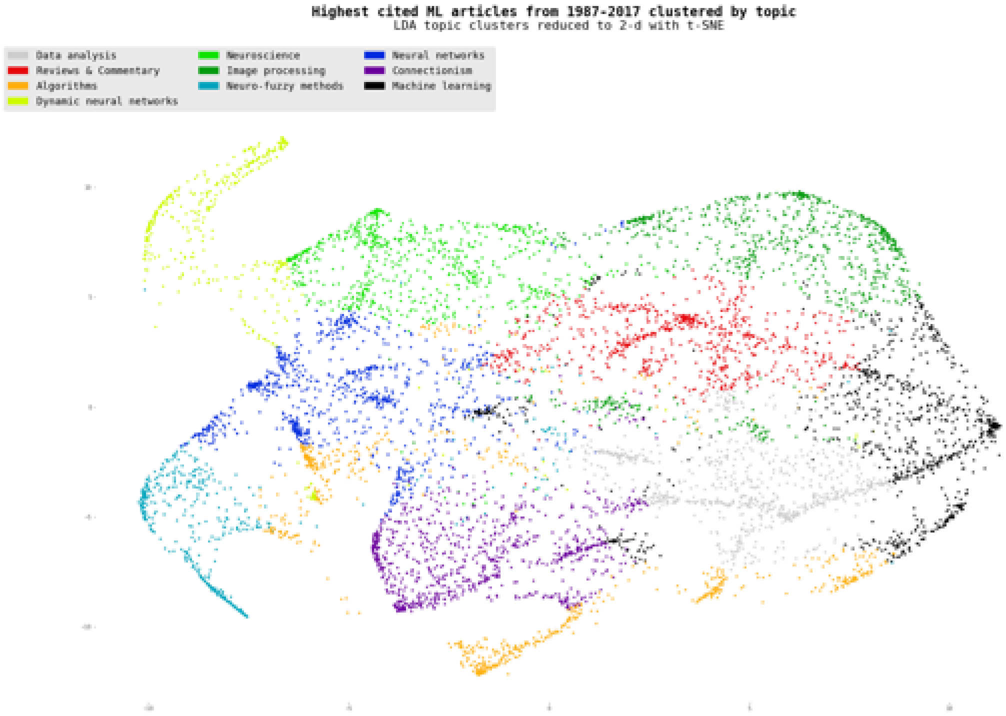


Figure D.3: LDA topics visualized with t-SNE for the Web of Science ML dataset.

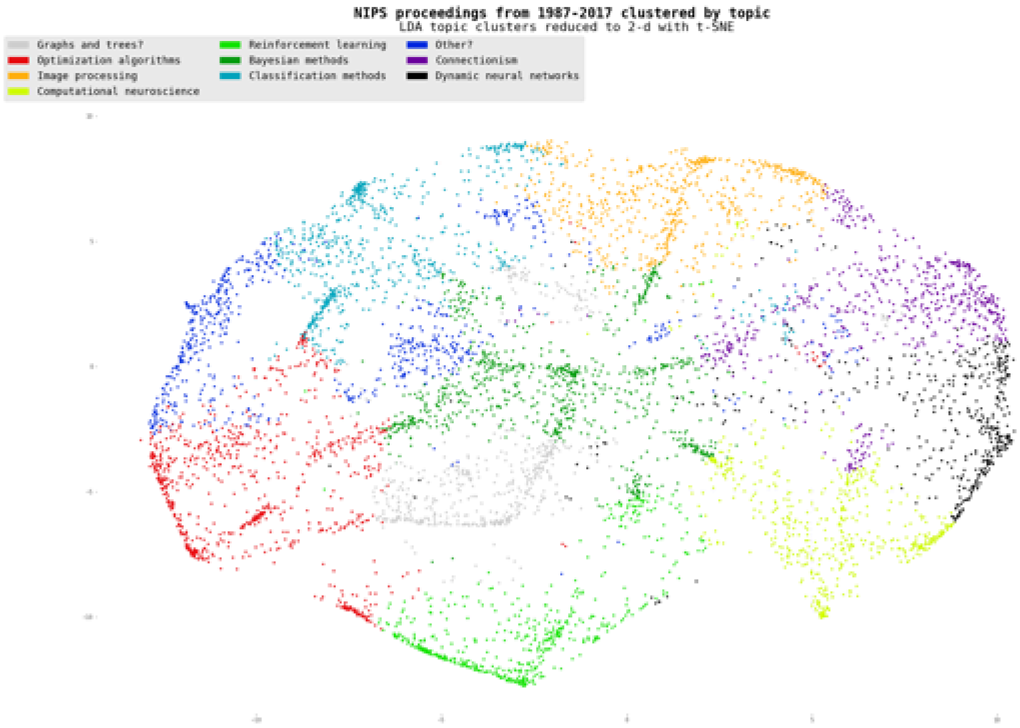
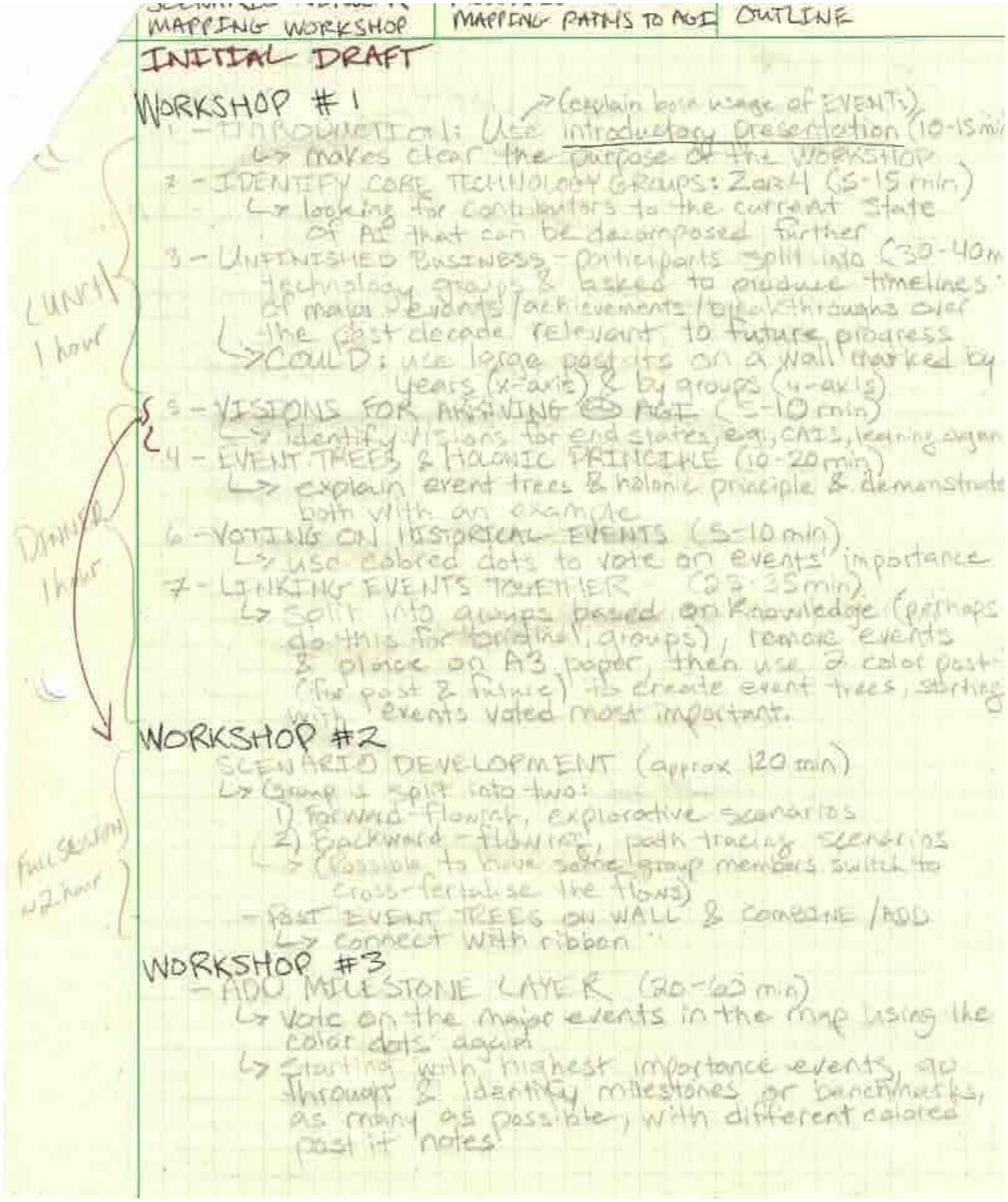


Figure D.4: LDA topics visualized with t-SNE for the NIPS dataset.

APPENDIX E: INITIAL SNM WORKSHOP PLANS

Each of the SNM workshops involved extensive preparation, after which a formal plan was made for each stage of the workshop. These plans are depicted here for each of the three workshops. These plans were all modified significantly due to numerous factors including circumstances encountered onsite at the workshop locations and the backgrounds of participants in the workshops who were not preselected. The plans depicted here are in the raw format directly as they were used for facilitation at the beginning of each round. Thus, some markup such as strikethrough is included – thus, this is not in error or a typo.

E1 First SNM Workshop Plan



MAPPING WORKSHOP | MAPPING PATHS TO AGI | OUTLINE

INITIAL DRAFT

WORKSHOP #1

1 – INTRODUCTION: Use introductory presentation (10-15 min) → (explain basic usage of EVENTs)

↳ makes clear the purpose of the WORKSHOP

2 – IDENTIFY CORE TECHNOLOGY GROUPS: 2 or 4 (5-15 min)

↳ looking for contributors to the current state of AI that can be decomposed further

3 – UNFINISHED BUSINESS – participants split into technology groups & asked to produce timelines of major events/achievements/breakthroughs over the past decade relevant to future progress (30-40 min)

↳ COULD: use large post-its on a wall marked by years (x-axis) & by groups (y-axis)

5 – VISIONS FOR ARRIVING AT AGI (5-10 min)

↳ identify visions for end states, e.g., CAIS, learning organ[illegible]

4 – EVENT TREES & HOLONIC PRINCIPLE (10-20 min)

↳ explain event trees & holonic principle & demonstrate both with an example

6 – VOTING ON HISTORICAL EVENTS (5-10 min)

↳ use colored dots to vote on events' importance

7 – LINKING EVENTS TOGETHER (25-35 min)

↳ split into groups based on knowledge (perhaps do this for original groups), remove events & place on A3 paper, then use 2 color post-its (for past & future) to create event trees, starting with events voted most important.

LUNCH 1 hour

DINNER 1 hour

WORKSHOP #2

SCENARIO DEVELOPMENT (approx 120 min)

↳ Group is split into two:

1) Forward-flowing, explorative scenarios

2) Backward-flowing, path tracing scenarios

↳ (Possible to have some group members switch to cross-fertalise the flows)

– POST EVENT TREES ON WALL & COMBINE/ADD

↳ connect with ribbon

FULL SESSION ~2 hour

WORKSHOP #3

– ADD MILESTONE LAYER (20-60 min)

↳ Vote on the major events in the map using the color dots again

↳ Starting with highest importance events, go through & identify milestones or benchmarks, as many as possible, with different colored post it notes

Figure E1.1: Preparation notes for the first SNM workshop in Spain. There were some adjustments made that involved the combination of workshops 2 & 3 into a single workshop due to scheduling constraints.

E2 Second SNM Workshop Plan

Session 1 (AM, day one)

- Introduction
  - Explain the purpose of the workshopping process (and the use as a developmental workshop)
  - Use whiteboard for simple demonstrations of the map that we're trying to develop
- Split into two groups
  - Group A
    - Works to identify items comprising a current research agenda for the broader scope of AI research
    - Works to identify event tree trunks for technologies that will arise with progress on the research agenda
  - Group B
    - Works to identify components of desirable intelligence that would be expected to comprise a transformative AI (TAI)
    - Works to identify event tree trunks for technologies that will enable these elements of intelligence comprising a TAI
- Place trunks on wall

Session 2 (PM [after lunch], day one)

- As a group discuss consequences
  - Identify end consequences and intermediate consequences
    - Add end consequences to wall
    - Add intermediate consequences to whiteboard
- Example event tree
- Split into same two groups
  - Groups A & B
    - Each group splits further into subgroups creates event trees for each of the trunks that were identified in the previous session
- Disperse stickers for voting on the most beneficial, both beneficial and dangerous, and most dangerous components (green, yellow, red)

In-between sessions: facilitators aggregate and condense event trees as best as possible.

Session 3 (day two)[164]

- Facilitators lead group to reduce and combine event trees
- Facilitators lead group to connect event trees with ribbon
- Work on identifying actions that can mitigate the risks/accelerate benefits present in the map, i.e. interventions

[164] This was not completed as planned.

    - Layer on top of map
- Facilitators lead group to identify effects of interventions on map and use alternate colored ribbon to show this

E3 Third SNM Workshop Plan

Workshop 1 (~3 hours):

Introduction

- That it's all there is to it and why
- About time, breaks, food, etc.
- FAQ
- AI vs. AGI

Making a map of the past until now

1. Identify key technologies in AI today.
    - (Here we can build from the previous two SNM workshop sessions.)
2. Divide into groups of these technologies.

3G. (group). For each of the technologies, identify important key events over the past 10 years (you may have done so earlier, but the focus is on 2010->).

4G. Vote for the most important events.

5G. Create event trees based on the most important key events.

6. Throw all the trees in one block and place them in a clear position.

Result: A map of events from past to present with detailed key events

Determining the images of the future, where AGI is achieved

1. Each expert will take turns to offer his or her own version of the image of the future where AGI has been achieved. Write down the facilitators.
2. Vote for the most important ones.
3. Break up into groups on the most important ones.

4G. Consider these images of the future in more detail, think about alternative, possibly counterintuitive worlds.

Result: the most important images of the future where AGI is achieved, with thoughtful alternatives in different worlds (?).

- The most plausible unique scenarios for AGI development, comprehensive AI services, human-brain interface or global catastrophe (e.g. AGI dystopia or AGI with significant human suffering, wealth inequality, enslavement, etc.)
    - We may want to ignore human-brain interface and comprehensive AI services so our map doesn’t become too broad. AGI scenarios and global catastrophe are important because they are both potential outcomes for AGI development.

Workshop 2 (~5 hours)

Forward-flow from past to present, backward-flow from future (2 groups)

1. Split into two halves, form subgroups.

2.1G. One group makes a forward 'tree of events' from the 'map of the past', highlighting technologies that will be developed from the current day and will contribute to the creation of AGI.

2.1G.a First step involves

2.2G. Another group makes backward "event trees" from the images of the future, highlighting the technologies that make it possible to come to this future.

2.3G. The groups "change" with the results of their work, making non-destructive notes and additions

3BO (backoffice). Connect the trees into one, remove the duplicates. Probably done in backoffice by facilitators, after one of the workshop days.

4. Looking at the resulting tree, highlight the most important technologies.

5. Vote for prioritization.

6. Break up into groups. For each group, select the technologies that have been prioritized, taking into account the specific competencies of the group members.

7G. For these technologies, create event trees (midcasting).

8. Throw all the trees in one block and place them clearly.

Result: A map of events from the present to the future with detailed key technologies

Workshop 3 (~5 hours):

S1: Creating a consequences layer

1. Split into groups from previous day (assign new participants to a group)
2. Have groups trace

Result: Identify the consequences of technologies in the map

S2: Connecting maps of the past and the future

BO. Connect the trees of the past to the present and the present to the future in one.

1. Plug into the map you have received, think about it, write down your comments on the directions a) which elements are most important, b) where there are gaps, c) which elements need additional detail.
2. Vote for the most important elements of the map.
3. Joint compaction of the map by combining the elements, removing the unimportant one.
4. identify gaps and elements to be expanded.
5. Divide into groups by these spaces/elements.

6G. Build event trees for spaces/elements.

7. Build the resulting trees into the overall map.
8. Re-compress the map.

Result: A common map from the past to the future without unnecessary/small elements or elements that should be combined into one

Workshop 4 (3 hours):

S1: Adding a Milestone Layer:

1. Highlight key technology groups in the map.
2. Divide into groups of these technologies.

3G. Highlight the Mailstone groups on the way to developing this technology.

4. Vote for the most likely/important of them.
5. Remove the least likely/important ones.
6. Integrate the layer into the overall map.

Result: The shared map now has a layer of milestone elements

S2: Creation of travel scenarios

1. Different experts try to follow the arrows of the map, describing what is happening as a sequential path leading to the AGI. Facilitators portal (or other technical method?) this way as a separate scenario.
2. After the Expert Advisor has told us how he sees the way, the rest of us comment/add to it.
3. Repeat p.1-2 until there are at least as many travel scenarios as there are endpoints ("variants of the image of the future" from previous days), it is desirable to have more.

Result: a set of scenarios that can be told to someone based on the map with its elements and arrows

APPENDIX F: FULL DELPHI RESULTS

The results from the Delphi study were extensive due to the large variety of responses. These results were too lengthy to include in the primary text, so they are included here instead. The results of the first round of the Delphi are reported in the first section of this Appendix and the results of the second round, including the rankings, are included in the section of this Appendix.

F1 1st Round Results

F1.1 Questions

- What are the most important forecasting targets?
    - How do we define qualitative and quantitative measures of progress toward forecasting targets?
    - How can we decompose abstract AI technologies into more easily forecastable targets?
    - What questions/targets matter for practical, near-term decision making?
- What are the implications of timelines?
    - Should we focus on capabilities or the impact of AI systems?
    - How can forecasts be applied to identifying and mitigating risks?
- How do we best evaluate overall AI progress?

- What are the most useful indicators (e.g. compute, talent, investment/resources, economic impact, benchmark performance)?
    - What performance metrics are relevant and most effective?
        - How do we assess the quality of a metric/benchmark's signal?
        - Are existing (SOTA) benchmarks relevant or useful (i.e. strong signal)?
        - Should we focus on tasks or abilities for measuring and forecasting AI progress?
        - How would we develop a broader discipline for measuring and assessing progress in AI (like psychometrics)?
        - How do we best analyze/measure AI systems' abilities to generalize, understand language and perform common sense reasoning?

- How can we model AI progress?
    - What are the best methods for modeling given the correct variables?
    - Why is progress faster in some metrics than others?
    - Can independent variables be used to model AI progress effectively model progress in other fields/research domains?

- What are the most probable AI development scenarios?

    - How do we identify the most plausible paths for a variety of transformative AI technologies/systems?
    - What will be the new applications/services made possible by new AI technologies?
    - What impact does NLP have on AI capabilities?

- How do we produce the best forecasts?
    - How do we aggregate and report metrics?
    - What are/how do we develop the best qualitative/quantitative a priori models?

- How effective can long term forecasting of AI progress be?
    - How do we best validate forecasts of AI progress: historical data/near-term progress?

- How do we utilize forecasts to inform decision makers and to develop appropriate and measured initiatives/interventions?
    - Who are the relevant stakeholders/audiences for forecasts and how do we best report forecasts to each?
    - What are information hazards related to AI forecasts and how do we best make decisions about how to guard and disseminate forecasting data?
    - What can we learn from historical examples of policy making?

- How can we improve/make more useful conventions regarding forecasting questions and answers?
- How do we forecast the automatability of different types of unique human tasks?
- How can we collect data measuring human performance that can easily be compared to machine performance (e.g. next word prediction log loss)?
- Can we identify a minimum viable timeline (e.g. 10% of strong AI) for use by stakeholders and decision makers?
- What can we learn from existing long-range forecasting techniques (e.g. clionomics, K-wave theory, S-curves)?
- How do we best operationalize group forecasting efforts?
- How effective are existing methods at forecasting technology (e.g. prediction markets, the Delphi)?

F1.2 Methods

- Judgmental forecasting techniques:
    - Simulation & role-play games
    - Scenario analysis
    - Blue-team/red-team
    - Expert elicitation:
        - Delphi
        - Expert adjustment
    - Prediction markets
    - Forecasting tournaments
    - Calibration training

    - Aggregation of expert opinion
    - Immersive observation of AI labs
    - Identifying clear and effective forecasting targets
    - Conceptual progress acceleration survey (using pairwise comparisons)
- Statistical forecasting techniques:
    - Statistical modeling
        - Extrapolation
        - Bayesian methods
    - Benchmarks & metrics
        - Aggregating into metrics for human comparison
        - Item response theory
    - Data science (e.g. tech mining, bibliometrics, scientometrics)
        - Theoretical models
    - Machine learning modeling
    - Simulation
- Hybrid methods (i.e. statistical and judgmental)
- Other:
    - Probabilistic reasoning (e.g. the Doomsday argument)
    - In-depth analysis of specific questions
    - Literature review

F2 2nd Round Results

Table F2.1: Delphi 2nd Round Results – Importance & Feasibility of Questions

	importance	imputed importance	feasibility	imputed feasibility
How do we best validate forecasts of AI progress: historical data/near-term progress?	4.63	3.85	3.91	2.36
How do we utilize forecasts to inform decision makers and develop appropriate and measured initiatives/interventions?	4.54	4.10	3.79	3.58
What questions/targets matter for practical, near-term decision making?	4.48	4.02	3.94	3.62
What performance metrics are relevant and most effective?	4.26	3.73	3.70	3.62
How do we assess the quality of a metric/benchmark's signal?	4.20	3.75	3.62	3.36
How do we identify the most plausible paths for a variety of transformative AI technologies/systems?	4.20	4.03	3.38	2.86

What are information hazards related to AI forecasts and how do we best make decisions about how to guard and disseminate forecasting data?	4.19	3.73	3.90	3.89
How effective are existing methods at forecasting technology (e.g. prediction markets, the Delphi)?	4.14	3.60	3.89	3.31
How do we forecast the automatability of different types of unique human tasks?	4.13	3.32	3.41	3.31
How can we decompose abstract AI technologies into more easily forecastable targets?	4.12	3.82	3.75	3.74
How do we best evaluate overall AI progress?	4.12	3.91	3.30	2.93
What are the most probable AI development scenarios?	4.11	3.28	3.11	3.07
What are the most useful indicators (e.g. compute, talent, investment/resources, economic impact, benchmark performance)?	4.10	3.86	3.92	2.93
How effective can long term forecasting of AI progress be?	4.10	3.69	2.86	2.86
How do we aggregate and report metrics?	4.09	3.81	4.09	3.07
How can forecasts be applied to identifying and mitigating risks?	4.04	3.31	3.14	2.76
How do we produce the best forecasts?	4.03	3.49	3.27	2.94
How do we define qualitative and quantitative measures of progress toward forecasting targets?	4.00	3.39	3.58	3.05
What are/how do we develop the best qualitative/quantitative a priori models?	3.93	3.75	3.21	3.97
Are existing (SOTA) benchmarks relevant or useful (i.e. strong signal)?	3.85	3.94	3.87	3.21
How do we best operationalize group forecasting efforts?	3.79	3.63	3.63	3.80
Who are the relevant stakeholders/audiences for forecasts and how do we best report forecasts to each?	3.78	4.12	4.16	3.26
How do we best analyze/measure AI systems’ abilities to generalize, understand language and perform common sense reasoning?	3.77	2.88	2.89	2.90
How can we model AI progress?	3.76	3.62	3.41	2.78

What can we learn from existing long-range forecasting techniques (e.g. clionomics, K-wave theory, S-curves)?	3.71	3.25	3.81	2.73
What are the implications of timelines?	3.63	4.12	3.65	3.73
How can we collect data measuring human performance that can easily be compared to machine performance (e.g. next word prediction log loss)?	3.52	3.75	3.96	3.26
What can we learn from historical examples of policy making?	3.51	4.12	3.85	3.54
How can we improve/make more useful conventions regarding forecasting questions and answers?	3.44	3.49	3.41	3.30
Should we focus on capabilities or the impact of AI systems?	3.42	3.72	3.13	3.03
What will be the new applications/services made possible by new AI technologies?	3.41	4.05	3.16	2.82
Why is progress faster in some metrics than others?	3.40	3.35	3.47	3.04
What impact does NLP have on AI capabilities?	3.34	3.40	3.30	2.60
Can independent variables be used to model AI progress effectively model progress in other fields/research domains?	3.31	3.44	3.28	3.38
What are the best methods for modeling given the correct variables?	3.29	3.64	3.45	3.10
Should we focus on tasks or abilities for measuring and forecasting AI progress?	3.23	3.63	2.92	3.43
Can we identify a minimum viable timeline (e.g. 10% of strong AI) for use by stakeholders and decision makers?	3.14	3.55	2.89	3.69
How would we develop a broader discipline for measuring and assessing progress in AI (like psychometrics)?	3.00	2.97	3.04	2.46

Table F2.2: Delphi 2nd Round Results – Importance & Feasibility of Methods

	importance	imputed importance	feasibility	imputed feasibility
Hybrid methods (i.e. statistical and judgmental)	4.52	3.41	3.66	3.38
Identifying clear and effective forecasting targets	4.29	3.71	3.75	3.12
Aggregating into metrics for human comparison	4.20	3.58	3.65	3.57
In-depth analysis of specific questions	4.19	3.34	3.86	2.93
Extrapolation	4.07	3.66	4.02	3.41
Theoretical models	3.98	3.47	3.42	3.58
Statistical modeling	3.90	3.48	3.69	3.07
Bayesian methods	3.88	3.74	3.89	3.82
Literature review	3.88	4.07	4.25	3.56
Immersive observation of AI labs	3.84	3.23	3.17	3.79
Machine learning modeling	3.83	3.67	3.69	3.08
Benchmarks & metrics	3.82	3.66	3.84	3.58
Blue-team/red-team	3.80	3.51	3.91	3.72
Expert adjustment	3.80	3.36	3.91	3.81
Conceptual progress acceleration survey (using pairwise comparisons)	3.78	3.87	3.81	3.44
Delphi	3.69	3.42	3.87	3.19
Data science (e.g. tech mining, bibliometrics, scientometrics)	3.67	3.33	3.64	3.34
Scenario analysis	3.59	3.05	3.62	3.91
Simulation	3.59	3.59	3.55	3.47
Statistical forecasting techniques:	3.50	3.53	3.41	3.33
Expert elicitation:	3.38	3.72	3.69	3.67
Aggregation of expert opinion	3.34	3.02	3.64	3.45
Probabilistic reasoning (e.g. the Doomsday argument)	3.25	3.90	3.43	3.47
Prediction markets	3.18	3.50	3.58	3.90
Judgmental forecasting techniques:	3.00	4.01	3.77	3.90
Item response theory	3.00	3.71	3.83	3.41
Forecasting tournaments	2.99	3.18	3.50	3.46
Simulation & role-play games	2.91	3.39	3.88	3.14
Calibration training	2.79	3.14	3.48	3.42

APPENDIX G: THE DELPHI RESEARCH AGENDA

The Delphi results were broken down to 3 groups of clusters for questions and 3 different classes of methods. Each high-level item in this research agenda corresponds to either a cluster or class of methods. *More details regarding the research agenda can be found in Gruetzemacher et al. (2020).* The research agenda itself, and the results of the Delphi – other than facilitator notes – were beyond the scope of this study. The purpose of the Delphi was to demonstrate and evaluate its utility for identifying and ranking salient research questions or forecasting targets. The research agenda is simply included here as a demonstration of the effectiveness of the process.

I. Many of the topics presented in this subsection could benefit from literature reviews which attempt to identify previous work that is relevant in any way.

II. Decomposition of forecasting targets is widely used in the presence of high uncertainty (MacGregor 2001), and it would be useful to demonstrate steps for effectively using this technique in the context of AI.

   A. This could involve an experiment to forecast benchmark performance on some measure (i.e. SuperGLUE (Wang et al. 2019)) at a given time (e.g. January 1st, 2021) using a model of two input indicators such as largest trained model size (in parameters) and largest cleaned dataset size (in GB). This would be a relatively simple experiment to carry out. (This would be similar to recent work by Kaplan et al. 2020, but for forecasting.)

B. Another way to validate the use of this technique in context would be to look at historical data for benchmarks, and identify one or more indicators that could be used as input to a model which could effectively forecast the benchmark of interest.

III. Several of the questions mentioned in this subsection could also benefit from expert elicitation. The Delphi technique, interviews or surveys could shed light on:

A. What capabilities or impacts would be more useful for decision makers?

B. What questions would be most beneficial for improving practical, near term decision making.

IV. Again, the foremost concrete suggestion is to conduct reviews of the existing literature.

A. Benchmarks and measures of progress have been successful for forecasting an AI milestone in at least one case (Russel and Norvig 1995), however, more recently they have been performed poorly at measuring progress in very popular research areas (e.g. GLUE and SuperGLUE; Wang et al. 2018, Wang et al. 2019).

1. A broad literature review of historical narrow AI benchmarks as well as more recent narrow AI benchmarks could be very valuable.
2. This could also be particularly useful for starting on the questions that are relevant in broader forecasting contexts (e.g. Q7a and Q8a).

V. Identifying the most plausible paths toward a variety of transformative AI systems is a topic that has received little attention, however, Gruetzemacher (2019a) makes some suggestions for future work.

VI. The dissemination of forecasts is a tricky but crucial issue; one common technique is scenario planning (Roper et al. 2011), but there is no magic bullet.

A. Yet again, a good place to start would be a literature review on the topic which may shed light on how to proceed with respect to determining the best approaches for delivering complex AI forecasts to policy makers and decision makers.

VII. Future of work research is an important topic which is receiving substantial attention already involving both data-based methods (Das et al. 2020; Martinez-Plumed et al. 2020) and expert elicitation (Duckworth et al. 2019).

   A. The data-based techniques explored here only scratch the surface; efforts to obtain more datasets and to combine them with disparate data sources, either public or private, are worthwhile.

      1. Due to this being a new field of study, a literature review of existing tech mining techniques may illuminate research paths not yet explored.

   B. Existing work on this topic has suggested that it is possible to automate close to 50% of human jobs in coming decades (Frey and Osborne 2017), however, little work has explored the potential for extreme labor displacement from AI (Gruetzemacher et al. 2019a).

      1. Models that can account for discontinuous progress in narrow domains (or more broadly) could be very useful for helping policy makers and organizations prepare for unforeseen scenarios. Thus, research on this topic can be very useful.

VIII. Extrapolation is the simplest forecasting technique yet it remains one of the most valuable, even for the purpose of forecasting AI progress. The challenge lies not in extrapolating a trend from data, but from identifying an indicator with sufficient data that is also a signal of something important to decision makers.

   A. Thus, we always recommend thinking critically about what may be a good indicator of AI progress. This does not necessarily require focus and dedicated time, but rather motivation and genuine interest in understanding AI progress. (For this reason, for some people this is a low hanging fruit).

IX. In the previous section, one question suggested that it is important to further explore the existing forecasting techniques to determine their relative effectiveness for 1) technological forecasting and 2) AI forecasting. This is necessary because little work has been done on such comparisons. Examples include:

A. Conducting a Delphi study involving PhD students studying AI using the same forecasting targets as those posted to ai.metaculus. (We feel this is a low hanging fruit.)

B. Conduct a survey and structured interviews with PhD students studying AI using the same forecasting targets as those posted to ai.metaculus (on or near the closing date for the forecasts). (We feel this is a low hanging fruit.)

X. With PhD students studying AI, use an established method to conduct scenario analysis on near-term plausible forecasting targets such as facial recognition technology, autonomous vehicles and lethal autonomous weapons. Perform this process for multiple groups of students, focusing on a one- or two-year time horizon and meticulously documenting the facilitation process. Evaluate the results to identify how the technique can be improved.

XI. Little work exists on hybrid methods, but Gruetzemacher (2019a) is a good starting point for interested researchers. Due to the multitude of forms this could take we do not elaborate here.

XII. In-depth analysis of specific questions casts a very broad net for possible research topics, and we hope that many readers are able to do better than us. However, examples include:

A. What indicators or milestones could be expected to precede discontinuous progress toward radically transformative AI?

B. Would a complete solution to the problem of meta-learning, in combination with a suite of powerful, specialized deep learning subsystems, be enough to enable some form of radically transformative AI?

APPENDIX H: INSTITUTIONAL REVIEW BOARD APPROVAL

Institutional Review Board (IRB) approval for evaluation of techniques involving human subjects was obtained for each technique that was evaluated in the study involving human subjects. These components of the study include (in chronological order): the practitioner survey, the structured interviews, the judgmental distillation mapping interviews, the scenario network mapping workshops and the Delphi process. The IRB protocol number at Auburn University was 17-484 EX 1804.

The approval from the IRB for these different components spanned nearly two years while the data for each component of the study was collected. The same IRB Information Letter format was used for all of the studies, with minor modifications for each of the different components. On the unnumbered pages that follow the following original, stamped IRB documents are included:

1. The initial Information Letter for the practitioner survey. This was used for both the survey and the structured interviews in Stockholm and Prague. (pp. 308).
2. The following form was for Consent to be Quoted. This was used for interview participants who were well-known and who were asked for consent to be quoted in this study or any published work resulting from it. For the purpose of the interviews, this form was printed directly on the back of the Information Letter being used at the time of the interviews. (pp. 309).
3. This is an email template for inviting experts to participate in the Delphi process for generating a research agenda. Similar email invite templates were used for inviting experts

to participate in interviews. However, modifications were made in order to tailor the invites to the experts. (pp. 310).

4. The final included document was the Information Letter at the end of the study, including the Delphi process, interviews and workshops. Different versions of the information letter were used at other points in the study, however, the primary components of all of these are contained in one of the two Information Letters presented here. (pp. 311).

AUBURN
UNIVERSITY

RAYMOND J. HARBERT
COLLEGE OF BUSINESS
DEPARTMENT OF SYSTEMS & TECHNOLOGY

(NOTE: DO NOT AGREE TO PARTICIPATE UNLESS AN IRB APPROVAL STAMP WITH CURRENT DATES HAS BEEN APPLIED TO THIS DOCUMENT.)

The Auburn University Institutional Review Board has approved this Document for use from 07/02/2018 to ---
Protocol # 17-484 EX 1804

INFORMATION LETTER
For a Research Study entitled
Confidence and Uncertainty in AI Predictions

You are invited to participate in a research study to forecast advancements and achievements in AI. This study is conducted by Ross Gruetzemacher, PhD candidate, under the direction of Dr. David Paradice, Harbert Eminent Scholar in the Auburn University College of Business, Department of Systems & Technology. You're invited to participate as an AI researcher age 19 or older.

What will be involved if you participate? If you decide to participate you will be asked to complete an interview. There are two options for the interview: a short format (~5 minutes) and a long format (~20 minutes).

Are there any risks or discomforts? No.

Will you receive compensation for participating? To thank you for your time you will be offered a $10 virtual Visa card for the short format interview or a $40 virtual Visa card for the long format interview. These options, for the respective formats, are close in value but not equal. They are made available for your convenience. You must complete the entire survey to receive a card.

If you change your mind about participating, you can withdraw at any time and your data will be deleted immediately.

Any data obtained in connection with this study will remain confidential unless you consent to be quoted (see back). Your voice will be recorded with a digital voice recorder for the survey and the data will be used later for qualitative and quantitative analysis. We will protect your privacy by encrypting the data and deleting it after publication of the study. Information collected will be analyzed qualitatively and quantitatively for forecasting purposes. It may be published or presented in academic research outlets.

If you have questions about this study, *please ask them now* or e-mail us later at rossg@auburn.edu or dparadice@auburn.edu.

If you have questions about your rights as a research participant, you may contact the Auburn University Office of Research Compliance or the Institutional Review Board by phone (334)-844-5966 or e-mail at IRBadmin@auburn.edu or IRBChair@auburn.edu.

HAVING READ THE INFORMATION PROVIDED, YOU MUST DECIDE IF YOU WANT TO PARTICIPATE IN THIS RESEARCH PROJECT. IF YOU DECIDE TO PARTICIPATE, THE DATA YOU PROVIDE WILL SERVE AS YOUR AGREEMENT TO DO SO. THIS LETTER IS YOURS TO KEEP.

____
Ross Gruetzemacher, Principal Investigator Date

____
Dr. David Paradice, Co-Investigator Date

403 LOWDER HALL
AUBURN, AL 36849-5247

TELEPHONE:
(334) 844-4908

ROSS GRUETZEMACHER
ROSSG@AUBURN.EDU

DR. DAVID PARADICE
DPARADICE@AUBURN.EDU

www.auburn.edu

CONSENT TO BE QUOTED

(NOTE: PLEASE DISREGARD THIS PAGE IF LEFT UNSIGNED.)

THIS IS AN ADDENDUM TO THE INFORMATION LETTER
For a Research Study entitled
Confidence and Uncertainty in AI Predictions

You have been asked for consent to quote your responses to the survey in future work to be presented and published as a result of this study. Only quotes relevant to the questions in the survey will be used.

BY SIGNING BELOW, YOU ARE CONSENTING TO YOUR RESPONSES FROM THE SURVEY BEING QUOTED.

____

Participant's signature   Date

Dear <expert>,

I am a graduate student in the Department of Systems and Technology at Auburn University. I would like to invite you to participate in my research study entitled Confidence and Uncertainty in AI predictions. You have been invited to participate in a Delphi study as an expert in the area of forecasting or AI.

If you choose to participate, you will be asked to participate in a Delphi study to collect data for a research agenda (~20 minutes to 2 hours). The study will take place beginning now and will last for four weeks. You will be able to participate at times that are convenient for you to respond to the questionnaire and to rank the results that are reported back to you after the first and second rounds. Upon completion of the study, for your participation, you will be compensated with a $40 Visa gift card.

Your responses will be made anonymous to the other participants, as is standard for Delphi studies. Only your email address will be retained in order to contact you to dispense your compensation. Your email address will be deleted after the study. If you would like to know more information about this study, an information letter can be found at this link. If you choose to participate after reading the information letter, you can begin the study from the link contained within.

If you have any questions, please contact me at rossg@auburn.edu or my advisor, Dr. Paradice, at dparadice@auburn.edu.

Thank you for your consideration,

Ross Gruetzemacher
PhD Candidate
Systems and Technology
Harbert College of Business
Auburn University

The Auburn University Institutional
Review Board has approved this
Document for use from
3/5/2020 to --------------
Protocol # 17-484 EX 1804

AUBURN
UNIVERSITY

RAYMOND J. HARBERT
COLLEGE OF BUSINESS
DEPARTMENT OF SYSTEMS & TECHNOLOGY

(NOTE: DO NOT AGREE TO PARTICIPATE UNLESS AN IRB APPROVAL STAMP WITH CURRENT DATES HAS BEEN APPLIED TO THIS DOCUMENT.)

INFORMATION LETTER

For a Research Study entitled

Confidence and Uncertainty in AI Predictions

You are invited to participate in a research study to forecast advancements and achievements in AI. This study is conducted by Ross Gruetzemacher, PhD candidate, under the direction of Dr. David Paradice, Harbert Eminent Scholar in the Auburn University College of Business, Department of Systems & Technology. You're invited to participate as an AI researcher age 19 or older.

What will be involved if you participate? If you decide to participate you will be asked to complete an interview (~20 minutes), engage in a Delphi study (~20 minutes to 2 hours; online questionnaire involving one or two rounds of follow-up questionnaires), or engage in a workshop (4-8 hours).

Are there any risks or discomforts? There is a risk of breach of confidentiality.

Will you receive compensation for participating? To thank you for your time you will be offered a $40 virtual Visa card. You must complete the entire interview to receive a card.

If you change your mind about participating, you can withdraw at any time and your data will be deleted immediately.

Any data obtained in connection with this study will remain confidential unless you consent to be quoted (see back). Your voice will be recorded with a digital voice recorder during the interview and the data will be used later for qualitative and quantitative analysis. We will protect your privacy by encrypting the data and deleting it after publication of the study. Information collected will be analyzed qualitatively and quantitatively for forecasting purposes. It may be published or presented in academic research outlets.

If you have questions about this study, *please ask them now* or e-mail us later at rossg@auburn.edu or dparadice@auburn.edu.

If you have questions about your rights as a research participant, you may contact the Auburn University Office of Research Compliance or the Institutional Review Board by phone (334)-844-5966 or e-mail at IRBadmin@auburn.edu or IRBChair@auburn.edu.

HAVING READ THE INFORMATION PROVIDED, YOU MUST DECIDE IF YOU WANT TO PARTICIPATE IN THIS RESEARCH PROJECT. IF YOU DECIDE TO PARTICIPATE, THE DATA YOU PROVIDE WILL SERVE AS YOUR AGREEMENT TO DO SO. THIS LETTER IS YOURS TO KEEP.

If you've been invited to participate in the Delphi study, please CLICK HERE.

Ross G 2/24/20

Ross Gruetzemacher, Principal Investigator Date

David Paradice 2/24/2020

Dr. David Paradice, Co-Investigator Date

403 LOWDER HALL

AUBURN, AL 36849-5247

TELEPHONE:

(334) 844-4908

ROSS GRUETZEMACHER
ROSSG@AUBURN.EDU

DR. DAVID PARADICE
DPARADICE@AUBURN.EDU

www.auburn.edu