big data and cognitive computing

Artificial Superintelligence

Coordination & Strategy

Edited by

Roman V. Yampolskiy and Allison Duettmann

Printed Edition of the Special Issue Published in *Big Data and Cognitive Computing*

www.mdpi.com/journal/BDCC

MDPI

Artificial Superintelligence

Artificial Superintelligence

Coordination & Strategy

Special Issue Editors

Roman V. Yampolskiy
Allison Duettmann

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade

Special Issue Editors

Roman V. Yampolskiy
University of Louisville
USA

Allison Duettmann
Foresight Institute
USA

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Actuators* (ISSN 2076-0825) in 2019 (available at: https://www.mdpi.com/journal/BDCC/special_issues/Artificial_Superintelligence).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03921-854-7 (Pbk)
ISBN 978-3-03921-855-4 (PDF)

Contents

About the Special Issue Editors

Roman V. Yampolskiy is a Tenured Associate Professor in the department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville. He is the founding and current director of the Cyber Security Lab and an author of many books including Artificial Superintelligence: a Futuristic Approach. During his tenure at UofL, Dr. Yampolskiy has been recognized as: Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a Senior member of IEEE and AGI; Member of Kentucky Academy of Science, and Research Associate of GCRI. Dr. Yampolskiy's main areas of interest are AI Safety and Cybersecurity. Dr. Yampolskiy is an author of over 100 publications including multiple journal articles and books. His research has been cited by 1000+ scientists and profiled in popular magazines both American and foreign (New Scientist, Poker Magazine, Science World Magazine), hundreds of websites (BBC, MSNBC, Yahoo! News), on radio (German National Radio, Swedish National Radio) and TV. Dr. Yampolskiy's research has been featured 700+ times in numerous media reports in 30 languages. Dr. Yampolskiy has been an invited speaker at 100+ events including Swedish National Academy of Science, Supreme Court of Korea, Princeton University and many others.

Allison Duettmann conducts research and coordinates Foresight Institute's technical programs. Her research focus is on the reduction of existential risks, especially from AI. At Existentialhope.com she keeps an index of readings, podcasts, organizations, and people that inspire an optimistic long-term vision for humanity. Allison speaks on existential risks & existential hope, AI safety, longevity, cryptocommerce, and other topics in ethics and technology. Prior engagements include Wall Street Journal, SXSW, The O'Reilly AI Conference, The World Economic Forum, The Partnership on AI, and Effective Altruism Global. Prior to Foresight, she hosted workshops, conferences and TEDx for corporations, governments, and the public across Europe, Latin America and the US. Allison holds an MS in Philosophy & Public Policy from the London School of Economics, where she developed an ethical framework for Artificial General Intelligence that uses NLP to aggregate crude ethical heuristics from texts.

Preface to "Artificial Superintelligence"

Focus of the AI safety community has increasingly started to include strategic considerations of coordination amongst relevant actors in the field of AI and AI safety, in addition to the steadily growing work on the technical considerations of building safe AI systems.

There are several reasons for this shift:

Multiplier Effects: Given the challenges of building safe AI systems (e.g., ethical, technical alignment, and cybersecurity concerns) we ought to ensure that the alotted timeframe is sufficient to develop thorough solutions. Coordination efforts could allow actors who develop AI to slow down when necessary, rather than engage in adversarial races, which may lead to corner-cutting on safety issues.

Pragmatism: While furthering the coordination of actors in the AI space is a complex challenge, coordination itself is not a novel problem. Many of the relevant actors ensuring that progress toward superintelligence remains beneficial to humanity are already known. There is already a promising research pool on coordination problems, as well as historic precursors of high-stake coordination problems which we have some familiarity and experience with, suggesting useful research directions for AI coordination.

Urgency: With race dynamics amongst major powers slowly emerging in AI and related fields, developing strategies for coordination is urgent. Currently, there is a window of opportunity to shape the nature of the relationships between current and future actors, to ensure a beneficial outcome for humanity.

Given the above benefits of coordination between those working on a path to safe superintelligence, this book surveys promising research in this emerging field regarding AI safety. On a meta-level, the hope is that this book can serve as a map to inform those working in the field of AI coordination of other promising efforts. Creating an informed and proactive research cohort would avoid Unliteralist's Curse scenarios, in which different efforts duplicate or, unbeknownst, counter, other promising efforts, and would open up avenues for collaboration as well, thereby more generally serving AI coordination research.

While this book focuses on AI safety coordination, coordination is important to most other known existential risks (e.g., biotechnology risks), and future human-made existential risks, some of which might still be unknown. Thus, while most coordination strategies in this book will be specific to superintelligence, we hope that some insights yield "collateral benefits" for the reduction of other existential risks, by creating an overall civilizational framework that increases in robustness, resiliency, and antifragility.

Roman V. Yampolskiy, Allison Duettmann
Special Issue Editors

big data and
cognitive computing

MDPI

Article

Future-Ready Strategic Oversight of Multiple Artificial Superintelligence-Enabled Adaptive Learning Systems via Human-Centric Explainable AI-Empowered Predictive Optimizations of Educational Outcomes

Meng-Leong HOW

National Institute of Education, Nanyang Technological University Singapore, Singapore 639798, Singapore; mengleong.how@nie.edu.sg

Received: 31 May 2019; Accepted: 23 July 2019; Published: 31 July 2019

Abstract: Artificial intelligence-enabled adaptive learning systems (AI-ALS) have been increasingly utilized in education. Schools are usually afforded the freedom to deploy the AI-ALS that they prefer. However, even before artificial intelligence autonomously develops into artificial superintelligence in the future, it would be remiss to entirely leave the students to the AI-ALS without any independent oversight of the potential issues. For example, if the students score well in formative assessments within the AI-ALS but subsequently perform badly in paper-based post-tests, or if the relentless algorithm of a particular AI-ALS is suspected of causing undue stress for the students, they should be addressed by educational stakeholders. Policy makers and educational stakeholders should collaborate to analyze the data from multiple AI-ALS deployed in different schools to achieve strategic oversight. The current paper provides exemplars to illustrate how this future-ready strategic oversight could be implemented using an artificial intelligence-based Bayesian network software to analyze the data from five dissimilar AI-ALS, each deployed in a different school. Besides using descriptive analytics to reveal potential issues experienced by students within each AI-ALS, this human-centric AI-empowered approach also enables explainable predictive analytics of the students' learning outcomes in paper-based summative assessments after training is completed in each AI-ALS.

Keywords: future-ready; strategic oversight; artificial superintelligence; artificial intelligence; forecasting AI behavior; predictive optimization; simulations; Bayesian networks; adaptive learning systems; pedagogical motif; explainable AI; AI Thinking; human-in-the-loop; human-centric reasoning; policy making on AI

1. Introduction

Artificial intelligence (AI) [1] refers to the ability of human-made systems to mimic rudimentary human thought. The term "artificial superintelligence" [2] goes beyond this primary ability of AI; it refers to the capability of human-made systems that can surpass humans. For example, they might even be able to rapidly discover hidden motifs or patterns in the data and then make predictions, while humans might find it very challenging to apperceive these hidden patterns within the mind, or perform similar feats at the speeds and performance levels that these systems can. To be clear, it could be argued that an AI system does not care about the need to prove to humans that it has achieved human-like consciousness (also referred to as the state of "singularity" or "artificial general intelligence") in order to be validated, certified, or given the stamp of approval by humans, so that it can properly be accorded a definitional label of its level of AI. There would probably be no notifications from AI systems the day they autonomously become self-aware, regardless of whether humans like it or not.

Meanwhile, in lieu of that fateful day, researchers have observed in studies that we already have artificial superintelligence working inconspicuously and tirelessly in our midst [3–5]. In the field of education, since the 1950s, AI deployed in the form of adaptive learning systems (ALS) [6,7], which are contemporary forms of intelligent tutoring systems (ITS) [8], have been utilized to assist teachers in the training of students [9]. Great strides have been made by researchers and commercial companies toward creating ALS that are powered by artificial intelligence, and perhaps, even superintelligence [2], in the sense that some of them have—dare I say—already surpassed the human teacher in terms of the ability to relentlessly perform the task of one-to-one tutoring, initiate progress checks, and conduct remediation. They can concurrently perform these tasks, perpetually to an unlimited number of students, round the clock, whenever and wherever the students choose to learn [10]. The developers of ALS and the researchers who field-test them have often lauded improvements in learning gains, and efficiencies of learning similar amounts of subject content in reduced amounts of time [11]. The primary function of an ALS is to educe (draw out) the learning abilities of the students by making them solve problems [12].

The advent of AI has enabled advanced developments of ALS. In recent years, an artificial intelligence-enabled adaptive learning system (AI-ALS) might utilize, for example, a variant of the AI-based Bayesian Knowledge Tracing (BKT) [13] algorithm, or some other proprietary algorithms formed from an ensemble of multiple AI-based methods to make "adjustments in an educational environment in order to accommodate individual differences" to provide a personalized learning experience for each student [14]. An example of a procedure that an AI-ALS might use to interact with the student is: (1) present the student with a topic or sub-topic to learn, (2) present the student with learning material that illustrate the concepts, (3) initiate a short progress check quiz of each sub-topic for the student. If the student could consecutively correctly answer a few questions, the AI-ALS would deem that the student has "passed" the learning objective for that topic or sub-topic (which will be indicated as "topic_passed" in the dataset). Otherwise, the student would be remediated by the AI-ALS until the learning outcome is achieved, and (4) finally, after the student has passed the progress check quiz, the AI-ALS would unlock more topics or sub-topics that are considered to be "ready for learning" by the student (that will be indicated as "topic_ready_for_learning" in the dataset). The AI-ALS is often used in conjunction with the flipped learning pedagogy [15], where the students are expected to log into the AI-ALS and learn as much as they can on their own at home. Subsequently, when they are in the classroom, the teacher can spend the precious class time more effectively by helping students to address any learning issues that they might have.

The current paper does not purport to be an empirical study of the effectiveness of any current AI-ALS. Rather, it proffers a future-ready human-in-the-loop [16] analytical framework that is based upon intuitive human-centric probabilistic reasoning, which could be used to characterize the "pedagogical motifs" [17] of any number of AI-ALS that may be deployed in the future. So long as the data from those systems are available to human analysts, this framework would still be useful for education stakeholders to gain an oversight of the "timbre" of multiple AI-ALS that are deployed in schools, even if those AI-ALS in the future are artificially superintelligent.

2. Research Problem and Initial Hypothetical Conjecture

2.1. Research Problem

In reality, the Department of Education of a city or a state or a country might choose not to implement a policy that compels all of the schools to use one single AI-ALS that is provided by one vendor. Presumably, the schools would also rather have the freedom to choose the AI-ALS that they prefer to deploy for their students. However, it would be remiss if the students were entirely left to the AI-ALS. For example, if the students do very well in the formative assessment tests in the AI-ALS, but perform badly in the paper-based post-test, or if the relentless testing-checking-remediating-testing algorithm of a particular AI-ALS is suspected to be causing too much stress for the students, it would be of concern to educational stakeholders. Currently, the AI-ALS products available in the educational industry have the ability to autonomously strive to make the student achieve mastery of the topics that they are required to learn. However, they are not yet fully equipped (e.g., with sensors or by other means) to take noncognitive factors (e.g., ability to manage stress, psychological well-being, motivation, level of engagement, etc.) of the students into consideration [18]. This is where a human-in-the-loop approach that is proffered by the current paper would play a vital role in bridging the gaps. It can be used to inform educational stakeholders in areas where the developers of the AI-ALS might have overlooked.

Coordination efforts between the educational stakeholders, such as policy makers, school leaders, and teachers, to assess the risks and safeguard the safety of students who are using the AI-ALS (in terms of noncognitive factors [19–23], such as, for example, the psychological well-being, or emotional intelligence to manage stress) are, undeniably, of paramount importance. Researchers, such as Manheim [24], Perry and Uuk [25], Turchin, Denkenberger, and Green [26], Umbrello [27], Watson [28], and by Ziesche and Yampolskiy [29], have made efforts to analyze the issues, values, and benefits of strategies and coordination in artificial superintelligence. Yet, in the field of education, there is still a dearth in the extant literature regarding the area of coordination and safety in artificial superintelligence [30]. From the perspective of education policy makers, it would be interesting to help to coordinate the analysis of data from multiple AI-ALS deployed in different schools, so they would be able to "see the big picture" and assess the potential issues to know whether each AI-ALS in the respective school is helping (or not helping) the students, and take further steps to address problems if necessary. Human teachers would be able to address the gaps in the students' learning process where the AI-ALS could not, and help to alleviate stressful situations for the students if they are uncomfortable using the AI-ALS.

In the field of education, the question "would it be possible to predict the conditions during the use of an educational intervention (e.g., an AI system) to enhance optimal student performance in the paper-based summative assessments?" might intrigue educational stakeholders, such as policy makers, parents, students, and educational researchers [31,32]. However, to the authors' knowledge, it is beyond the scope of consideration by the developers of the AI-ALS to predict how the students' scores within the AI-ALS could influence their learning outcomes in a summative assessment (e.g., a paper-based standardized test that all the students are required to take in the school) after their training has been completed in the AI-ALS. To achieve this predictive capability, it is imperative for the pedagogical "motif" or "timbre" or "disposition" of the AI-ALS to be known, as each of them would interact with students in different ways. Although educational stakeholders need to examine the pedagogical characteristics of the AI-ALS, the vendors of the systems would understandably be reticent about divulging the exact algorithm to the customers, as they are closely-held trade secrets. Instead of believing all of the information provided by the vendors who are inclined to assure that everything will be excellent, it would be prudent for educational stakeholders to independently investigate the pedagogical characteristics that underlie these AI-ALS. Frameworks have been created by researchers for the evaluation of ALS [33]. Nevertheless, those laudable techniques were often formally presented as mathematical equations, which could prove to be difficult for educational stakeholders who might not have the necessary computer programming human resources or enough time to implement them. There remains a need for a more intuitive and practical way for educational stakeholders—rather than computer scientists—to apply human-in-the-loop AI-Thinking [34,35] and quickly achieve a strategic oversight of the multiple AI-ALS, which is crucial for informing educational policy and advancing pedagogical practice.

2.2. Initial Hypothetical Conjecture

The initial hypothetical conjecture assumes that the developers of an AI-ALS might have designed it to push the higher-performing students a little harder, and conversely, to go easy on the relatively lower-performing students. Therefore, it would not be unreasonable to imagine that a student who had performed poorly in the AI-ALS might have experienced having his or her weaknesses being educed (drawn out) by the system. Subsequently, after a personal reflection of those problems via vicarious trial and error (VTE) [36], the student could become cognizant of those weaknesses and could avoid similar predicaments during problem-solving in the paper-based post-test. Conversely, a student who had performed well in the AI-ALS might not have experienced having his or her weaknesses educed, and hence might lack the personal reflections or the VTE to learn from those experiences. Consequently, he or she might perform poorly in the post-test. The approach being proffered in the current paper would purely characterize its informational pattern (its motif), regardless of whether a student scored high or low within the AI-ALS. In other words, it does not affect the calculation of the "gains" that are attributed to the prowess of the AI-ALS, as it will not simply be a subtraction of the results of the paper-based post-test from the paper-based pre-test.

Nevertheless, it would be contrived to only measure the "gains" in terms of cognitive dimensions while using the pre-test and post-test, as there might be noncognitive benefits for the students too. Hence, a survey that could be used to understand more about the noncognitive aspects of their learning experiences could also be administered to the students upon the completion of their learning process in the AI-ALS. Some of the possible noncognitive instruments that could be utilized by educational stakeholders include those that are offered by researchers such as Al-Mutawah and Fateel [37], Chamberlin, Moore, and Parks [38], Egalite, Mills, and Greene [39], Lipnevich, MacCann, and Roberts [40], and Mantzicopoulos, Patrick, Strati, and Watson [41].

2.3. Potential Issues that Education Researchers Might Encounter

When a school decides to let a class of students use an AI-ALS to assist the teachers, it might not occur to the school leaders or teachers to make any arrangements for the formation of a control group. Understandably, the school may have concerns that parents might be unwilling to give permission for their children to participate in a control group, merely to form a baseline group for comparison with the treatment group, with no assistive benefits from any educational technology. Moreover, it will not be easy to perform direct comparisons between the treatment and control group even if a control group could be formed by the school, as the teaching experiences and skills of the teachers between the control and the treatment group might be unevenly matched. Further, it might not be surprising if some students from the treatment group or control group have the advantage of receiving extra help from tuition lessons outside of school. In effect, the myriad potential confounding factors would be difficult to account for, if fair comparisons must be performed between the treatment group that attended lessons where the teacher had been assisted by the AI-ALS to learn mathematics, and the control group that attended lessons where the teacher had not been assisted by the AI-ALS. Last but not least, a major problem that is faced by analysts who are considering the use of null hypothesis significance testing (NHST) frequentist approaches is that there might not be results that yield any meaningful statistical significant difference, due to the low number of participants in real-world situations (e.g., 20 students per class in each school) and the corresponding non-parametric data distributions [42].

Practical examples will be provided in the current paper to overcome these constraints. They will be used to illustrate how strategic oversight could be implemented using an artificial intelligence-based analytical tool by educational stakeholders to analyze data from five dissimilar AI-ALS deployed in small-scale pilot studies, each in a different school, and how conditions in those different AI-ALS could be used for predictive optimizations of educational outcomes in the paper-based summative assessments.

3. Methods

3.1. Rationale for Using the Bayesian Approach for Human-Centric Probabilistic Reasoning

Bayesian approaches for analyzing statistical data [43] have gained traction in behavioral science research in recent years [44]. The Bayesian network (BN) [45–47] approach is suitable for analyzing non-parametric data from a small number of participants, because it does not require the underlying variables of a model to assume or have a normal parametric distribution [42,48,49]. The Bayesian paradigm enables researchers to perform hypothesis testing by including prior knowledge into the analyses. Due to this capability, it becomes unnecessary to repeatedly perform multiple rounds of null hypothesis testing [50–52] when using Bayesian data analytical techniques.

Researchers in education, such as Kaplan [53], Levy [54], Mathys [55], and Muthén and Asparouhov [56], have employed the Bayesian approach to model the behavior of pedagogical systems operating under conditions with uncertainties, as the information about entropy in these systems could be harnessed to understand more about the factors that contribute (either positively or not) to their robustness and resiliency [57]. In educational technology, Bekele and McPherson [58] and Millán, Agosta, and Cruz [59] have also utilized the Bayesian approach, because it enables them to measure information gain, as depicted in Claude Shannon's Information Theory [60], which could be likened to the notion of learning by the students.

The primary advantage of BN is that its strong probabilistic theory empowers users to gain an intuitive understanding of the processes involved. It also enables predictive reasoning because, given observations of evidence, questions can be posed to find the posterior probability of any variable or set of variables. However, the current paper does not purport to perform comparisons between the use of BN and other AI-based techniques, such as artificial neural networks (ANN), as that has already been well-documented by Correa, Bielza, and Pamies-Teixeira [61]. They observe that BN can illustrate the relationships that exist between the nodes in a model to provide more information than an ANN, which has been likened to a black box.

3.2. The Bayesian Theorem

A succinct introduction to the Bayesian theorem and BN will be presented here. However, readers who are interested to learn more about BN are encouraged to peruse the works of Cowell, Dawid, Lauritzen, and Spiegelhalter [62]; Jensen [63]; and, Korb & Nicholson [64].

The mathematical theorem (see Equation (1)) for human-centric probabilistic reasoning was developed by the mathematician and theologian, Reverend Thomas Bayes, but he passed away and the notes were left unpublished in his drawer. They were later found and published posthumously by his friend Richard Price in 1763 [43].

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \tag{1}$$

According to Equation (1), H represents a hypothesis and E represents a piece of evidence. *P(H|E)* is referred to as the conditional probability of the hypothesis H, which means the likelihood of H occurring given the condition that the evidence E is true. It is also referred to as the posterior probability, which means the probability of the hypothesis H being true after calculating how the evidence E influences the verity of the hypothesis H.

P(H) and *P(E)* represent the probabilities of the likelihood of the hypothesis H being true, and of the likelihood of the evidence E being true, independent of each other, and it is referred to as the prior or marginal probability—*P(H)* and *P(E)*, respectively. *P(E|H)* represents the conditional probability of the evidence E, that is, the likelihood of E being true, given the condition that the hypothesis H is true. Hence, the quotient *P(E|H)/P(E)* represents the support that the evidence E provides for the hypothesis H.

3.3. The Research Model

The primary goal of the current paper is to offer one out of myriad possible ways that analytical collaboration between educational stakeholders could be performed for evaluation of potential issues by simulating how much (or how little) the learning of mathematics can be improved for the students in five different schools, which used five dissimilar AI-ALS that were provided by five vendors. The probabilistic reasoning techniques used are based on BN. Within the BN, the concept of the Markov Blanket [65], in conjunction with Response Surface Methodology (RSM) [66–69], are utilized, as they are proven techniques for examining the optimization of the relations between the variables of theoretical constructs, even if they are not physically related.

The Bayesian approach has been chosen, because it is a methodology that has been used for modeling the performances and knowledge of students; in particular, by the developers of adaptive learning software applications, such as Collins, Greer, and Huang [70]; Conati, Gertner, VanLehn, and Druzdze [71]; Jameson [72]; and, VanLehn, Niu, Siler, and Gertner [73]. However, these published works were focused on the vantage points of the developers who were describing the advantages of their respective products.

In contrast, it would be quite difficult for end-users of any AI-ALS to understand more about the inner workings of the proprietary algorithms that power the interactions with the students. The current paper proffers an approach that enables educational stakeholders to use descriptive analytics as well as predictive simulations to analyze the data that could be procured from the learners' performance reports in the server of an AI-ALS. This allows for analyses which could include comparisons and evaluations of multiple AI-ALS. The intention is to inform the educational stakeholders in each respective school, so that their teachers can remediate and bridge the gaps for the students, in whichever topics that the AI-ALS could not do so.

In Sections 4.5 and 4.6, the detailed BN model of the students' knowledge will be presented. It can inform educational stakeholders about the specific mathematics topics that the students are ready to learn, and the topics that they have already passed. Due to the coordination efforts between educational stakeholders in the five schools, they may use the vital information depicted by the relations between the nodes/variables in the BN to provide remediation for the students who are struggling in their studies. Hence, they could achieve better learning outcomes and decrease the probability of the potential risks that usage of an AI-ALS might entail (e.g., the students experiencing undue stress).

The BN model in the current paper is machine-learned from data procured from the scores of a paper-based pre-test, the learning progress scores while the students were using the AI-ALS, the Likert-scale scores from a survey, as well as the scores from a paper-based post-test. The current paper analyzes the relations using the generated BN. The theoretical constructs within the BN include the paper-based pre-test, the mediator (which is the AI-ALS), the paper-based post-test, and the noncognitive constructs (e.g., motivation, engagement, interest, self-regulation, etc.) in the survey. When researchers and educational stakeholders evaluate an AI-ALS, an understanding of these relations is essential for determining whether the interventions would be beneficial to the students. Therefore, the current paper proposes a practical Bayesian approach to demonstrate how educational stakeholders—rather than computer scientists—could analyze data from a small number of students. In order to explore the pedagogical motif of the AI-ALS, the following two types of analytics will be subsequently presented in Sections 4 and 5:

Descriptive analytics of "what has already happened?" in Section 4:

> Purpose: to use descriptive analytics to discover the pedagogical motifs of the five AI-ALS deployed in five different schools. For descriptive analytics, BN modeling in Section 4.7 will first utilize the parameter estimation algorithm to automatically detect the data distribution of each column in the dataset. Further descriptive statistical techniques that will be employed to understand more about the current baseline conditions of the students include quadrant analysis, curves analysis, and Pearson correlation analysis.

"What-if?" predictive analytics in Section 5:

> Purpose: to use predictive analytics to perform in-silico experiments with fully controllable parameters from the pre-test to the mediating intervention to the post-test for prediction of future outcomes. Instead of just simply measuring gains by subtracting the students' post-test scores from the pre-test scores, a probabilistic Bayesian approach

will be used to simulate counterfactual scenarios to better inform educators and policy makers about the pedagogical characteristics of the five AI-ALS that are being deployed in five different schools. For predictive analytics, counterfactual simulations in Section 5 will be employed to explore the pedagogical motif of the AI-ALS. In Section 6, the predictive performance of the BN model will be evaluated using tools that include the gains curve, the lift curve, the Receiver Operating Characteristic (ROC) curve, as well as by statistical bootstrapping of the data inside each column of the dataset (which is also the data distribution in each node of the BN model) by 100,000 times to generate a larger dataset to measure its precision, reliability, Gini index, lift index, calibration index, the binary log-loss, the correlation coefficient R, the coefficient of determination R2, root mean square error (RSME), and normalized root mean square error (NRSME).

4. Descriptive Analytics of "What has Already Happened?"

In this section, the procedures that were carried out in descriptive analytics to make sense of "what has already happened?" in the collected dataset will be presented. The dataset comprising 100 students (20 students from each school, from five different schools, all of whom were about 13–14 years old) who had used the AI-ALS, was imported into Bayesialab to deliberately illustrate the capabilities of BN in handling nonparametric statistical data from a small number of participants [74]. The purpose is to discover the informational "pedagogical motif" of the learning intervention generated by each AI-ALS. In the context of this study, the notion of "pedagogical motif" is conceptually defined as the pattern, timbre, disposition, and the unique characteristics with which each AI-ALS pedagogically interacts with the students.

4.1. The Dataset Procured from the Reports Generated by AI-ALS

The zip file containing the following datasets can be downloaded from https://doi.org/10.6084/m9.figshare.8206976.

The file "data_five_classes_AI_ALS.csv" contains the combined data of the five datasets from five different groups of students in different schools. For the convenience of the reader who may wish to import the data files from each group of students in each of the respective school into Bayesialab when prompted to do so in this paper, these files "data_ai_als_class_1.csv", "data_ai_als_class_2.csv", "data_ai_als_class_3.csv", "data_ai_als_class_4.csv", and "data_ai_als_class_5.csv" are also separately available in the zip file. The codebook describing the data, "ai-als-data_codebook.txt" is also included.

4.2. Codebook of the Dataset

The dataset could be procured from the reports that were generated by the server of each AI-ALS. Even though the variables from different datasets of the various AI-ALS would presumably be dissimilar, they could be aggregated to a form that is based on the mathematics topics and sub-topics (see Table A1 in Appendix A) that the students are required to learn in their curriculum. Each column in the dataset is presented as a node in the BN. It can be assumed that higher values in the data of both "math_topic_passed" (appended with the letter "P") and "math_topic_ready_for_learning" (appended with the letters "RL") are considered to be indicators of better performance, and vice-versa.

4.3. Software Used: Bayesialab

The software which will be utilized is Bayesialab [75]. A suggested pre-requisite activity for the reader is to peruse the free user-guide from http://www.bayesia.com/book/ before proceeding with the exemplars illustrated in the following sections, as it documents the tools and functionalities of the Bayesialab software.

4.4. Pre-Processing: Checking for Missing Values or Errors in the Data

It would be prudent to check the data (using the file "data_five_classes_AI_ALS.csv") for any anomalies or missing values before using Bayesialab to construct the BN. In the dataset used in this study, there were no anomalies or missing values. However, should other analysts encounter missing values in their datasets, they could use Bayesialab to predict and fill in those missing values, rather than discarding the row of data with a missing value. Bayesialab would be able to perform this by machine-learning the overall structural characteristics of that entire dataset being studied, before producing the predicted values. Bayesialab uses the Structural Expectation Maximization (EM) algorithms and Dynamic Imputation algorithms to calculate any missing values [76].

4.5. Overview of the BN Model

BN, which is also referred to as Belief Networks, Causal Probabilistic Networks, and Probabilistic Influence Diagrams are graphical models, which consist of nodes (variables) and arcs or arrows. Each node contains the data distribution of the respective variable. The arcs or arrows between the nodes represent the probabilities of the correlations between the variables [77].

Using BN, it becomes possible to use descriptive analytics to analyze the relations between the nodes (variables) and the manner in which initial probabilities, such as the number of hours spent in the AI-ALS and/or topics passed/ready to learn, and/or noncognitive factors, might influence the probabilities of future outcomes, such as the predicted learning performance of the students in the paper-based post-test.

Further, BN can also be used to perform counterfactual speculations regarding the initial states of the data distribution in the nodes (variables), given the final outcome. In the context of the current paper, exemplars will be presented in the predictive analytics segment (in Section 5) to illustrate how counterfactual simulations can be implemented while using BN. For example, we can simulate these hypothetical scenarios in the BN if we wish to find out the conditions of the initial states in the nodes (variables) that would lead to high probability of attaining high-level scores in the post-test, or if we wish to find out how to prevent students from attaining low scores or failing in the paper-based post-test.

The relation between each pair of connected nodes (variables) is determined by their respective Conditional Probability Table (CPT), which represents the probabilities of correlations between the data distributions of the parent node and the child node [78]. In the current paper, the values in the CPT are automatically machine-learned by Bayesialab, according to the data distribution of each column/variable/node in the dataset. Nevertheless, it is possible, but optional, for the user to manually enter the probability values into the CPT, if the human user wishes to override the machine learning software. In Bayesialab, the CPT of any node can be seen by double-clicking on it.

The BN model can be used to depict the data distribution of the students' score clusters (see Figure 1) in the AI-ALS in terms of the mathematics topics which include Arithmetic Readiness, Real Numbers, Linear Equations, Linear Inequalities, Functions and Lines, Exponents and Exponential Functions, Polynomials and Factoring, as well as Quadratic Functions and Equations. These score clusters were generated via machine-learning by the Bayesialab software. By generating this model from the data that contained varying levels of performance of the students (even if it was just 20 students from each school, with a total of 100 students from five schools), we could obtain a "pedagogical motif" of each AI-ALS, which meant that we could then perform simulations in each computational model to study how it could behave under certain conditions. This will be elaborated and presented later in Section 5.

4.6. Detailed Descriptions of the BN in the Current Paper

Nodes (both the blue round dots, as well as the round cornered rectangles showing the data distribution histograms) represent the variables of interest, for example, the score of a particular mathematics topic (connected to nodes with scores from their corresponding sub-topics), the number of hours that are spent by a student in the AI-ALS, the percentage of mathematics topics which a student had passed in the AI-ALS, or the rating of a particular noncognitive factor (e.g., motivation of a student). Such nodes can correspond to symbolic/categorical variables, numerical variables with discrete values, or discretized continuous variables. We exclusively discuss BN with discrete nodes in the current paper even though BN can handle continuous variables, as it is more relevant in helping educational stakeholders categorize students into high, mid, and low achievement groups, so that teachers can utilize differentiated methods to better address the students' learning needs.

Directed links (the arrows) could represent informational (statistical) or causal dependencies among the variables. The directions are used to define kinship relations, i.e., parent-child relationships. For example, X is the parent node of Y, and Y is the child node in a Bayesian network with a link from X to Y. In the current paper, it is important to note that the Bayesian network presented is the machine-learned result of probabilistic structural equation modeling (PSEM); the arrows represent the probabilistic structural relationships between the parent node and the child nodes. The first letter of the name of each node/data entity is presented in the upper case for better readability.

In the BN model used in the current paper (see Figure 1), the node representing the Pre-test results (from a paper-based math test) is connected to the "mediator" node representing the pedagogical motif of the AI-ALS, and subsequently the "mediator" node that represents the pedagogical motif of the AI-ALS is also connected to the node that represents the Post-test results (from another paper-based math test). This enables the probabilities of the AI-ALS as a mediator of the students' performance to be calculated, and subsequently it will be possible to simulate hypothetical scenarios (to be presented later in Section 5).

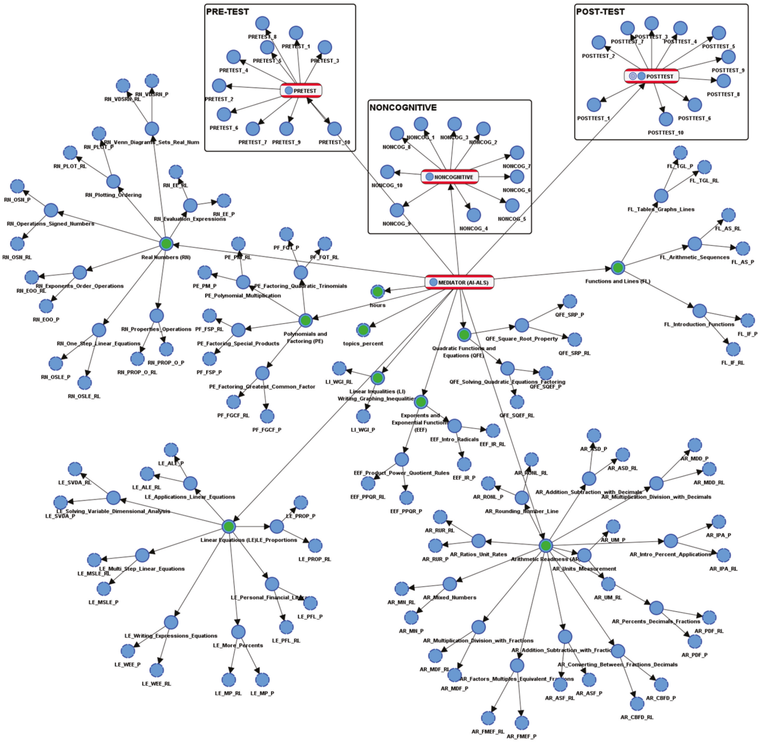


Figure 1. Full view of the Bayesian network: the component nodes (in blue) and the superordinate factor nodes (in green) were used for machine learning the overall performance of 100 students who had used the five different artificial intelligence-enabled adaptive learning systems (AI-ALS).

4.7. Descriptive Statistical Analysis of the Dataset

From the combined dataset of all the 100 students' performance who had used the five different AI-ALS (using the file "data_five_classes_AI_ALS.csv"), the following score-clusters machine-learned by Bayesialab were observed (see Figure 2):

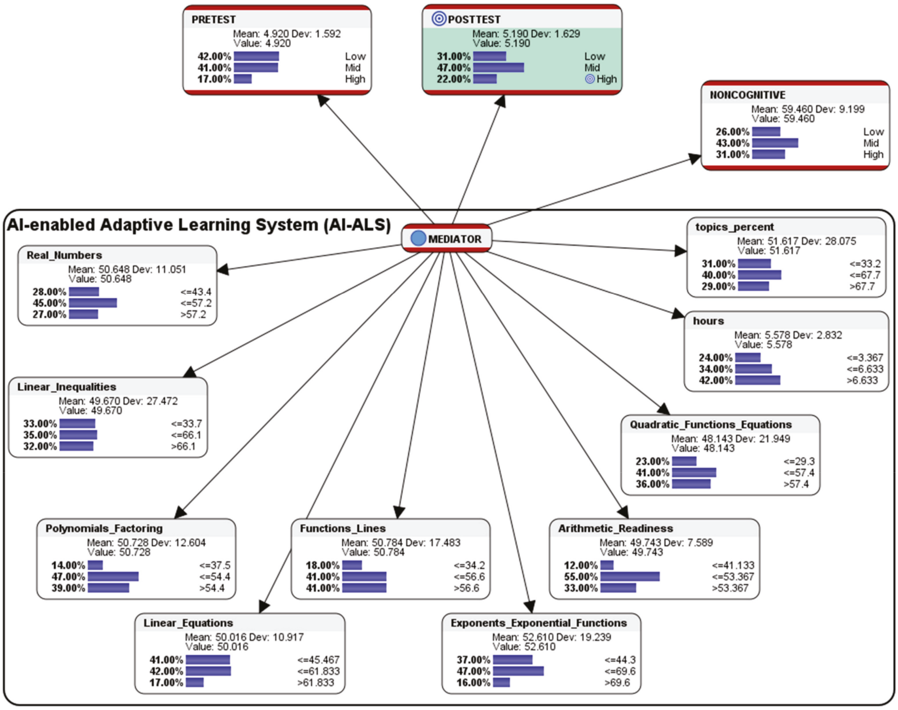


Figure 2. Simplified aggregated view of the Bayesian network previously shown in Figure 1, presenting only the superordinate factor nodes with their machine-learned score-clusters, depicting the overall performance levels of all 100 students who had used the five dissimilar AI-ALS from five different vendors.

In the paper-based Pre-test before the students used the AI-ALS, 42% of the students scored at the Low-level, 41% scored at the Mid-level, and 17% scored at the High-level. In the paper-based Post-test after the students had gone through the training within the AI-ALS, 31% scored at the Low-level, 47% scored at the Mid-level, and 22% scored at the High-level. Overall, in terms of conventional gains, there was an improvement of 11% of the students who had scored at the Low-level (a decrease from 42% in the Pre-test to 31% in the Post-test); there was an improvement of 6% in the students who had scored at the Mid-level (an increase from 41% in the Pre-test to 47% in the Post-test); and, there was an improvement of 5% in the students who had scored at the High-level (an increase from 17% in the Pre-test to 22% in the Post-test).

In the aggregated Noncognitive factor, 26% of the students were at the so-called Low-level, 43% were at the Mid-level, and 31% were at the High-level.

Within the AI-ALS, in the topic of Real Numbers, 28% of the students scored at the Low-level (<=43.4% of the total marks for Real Numbers), 45% scored at the Mid-level (>43.4 and <=57.2), and 27% scored at the High-level (>57.2).

In the topic of Linear Inequalities, 33% scored at the Low-level (<=33.7), 35% scored at the Mid-level (>33.7 and <=66.1), and 32% scored at the High-level (>66.1).

In the topic of Polynomials and Factoring, 14% of the students scored at the Low-level (<=37.5), 47% scored at the Mid-level (>37.5 and <=54.4), and 39% scored at the High-level (>54.4).

In the topic of Linear Equations, 41% of the students scored at the Low-level (<=45.467), 42% scored at the Mid-level (>45.467 and <=61.833), and 17% scored at the High-level (>61.833).

In the topic of Functions and Lines, 18% of the students scored at the Low-level (<=34.2), 41% scored at the Mid-level (>34.2 and <=56.5), and 41% scored at the High-level (>56.6).

In the topic of Exponents and Exponential Functions, 37% of the students scored at the Low-level (<=44.3), 47% scored at the Mid-level (>44.3 and <=69.6), and 16% scored at the High-level (>69.6).

In the topic of Arithmetic Readiness, 12% of the students scored at the Low-level (<=41.133), 55% scored at the Mid-level (>41.133 and <=53.367), and 33% scored at the High-level (>53.367).

In the topic of Quadratic Functions and Equations, 23% of the students scored at the Low-level (<=29.3), 41% scored at the Mid-level (>29.3 and <=57.4), and 36% scored at the High-level (>53.4).

Regarding the average number of hours spent by each student in the AI-ALS, 24% of the students were at the Low-level (<=3.367 h), 34% of the students were at the Mid-level (>3.367 and <=6.633 h), and 42% were at the High-level (>6.633 h).

In the percentage of the total number of topics that were mastered by the students in the AI-ALS, 31% of the students were at the Low-level (<=33.3%), 40% were at the Mid-level (>33.3% and <=67.7%), and 29% were at the High-level (>67.7%).

4.7.1. Descriptive Analytics: Profile Analysis of Each AI-ALS

A strategic overview of how the students performed (see Figures 3 and 4) could be accomplished via profile analysis. This tool can be activated in Bayesialab via these steps: *Bayesialab (validation mode) > Visual > Segment > Profile*.

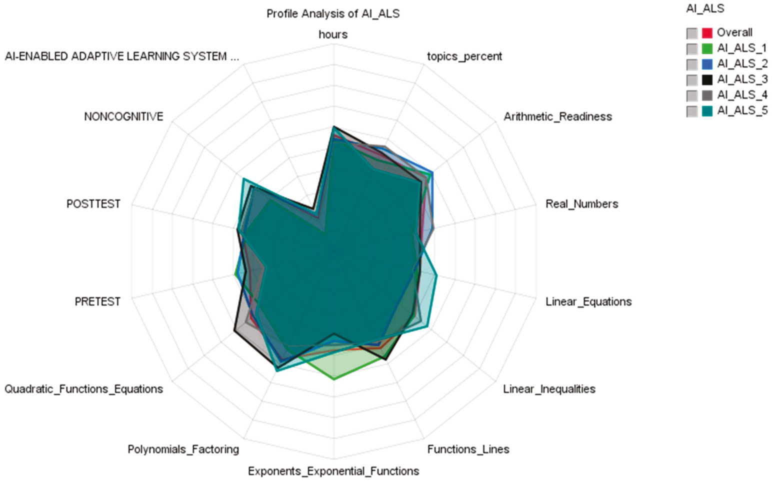


Figure 3. Profile analysis of the five groups of students, each of which had used a different AI-ALS.

Figure 4 is an alternative presentation of the profiles presenting the performance of the five groups of students in different schools, each of which had used a different AI-ALS.

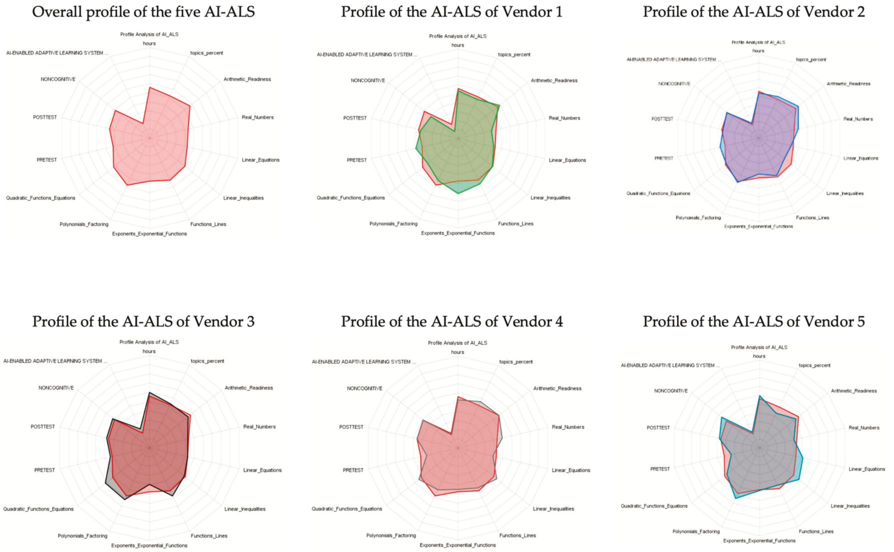


Figure 4. Profiles of five different AI-ALS, each from a different vendor, superimposed on top of the overall profile.

4.7.2. Descriptive Analytics: Quadrant Analysis

Comparison of Total Effects of the five different AI-ALS on the paper-based Post-test can be performed while using quadrant analysis. This tool can be activated in Bayesialab via these steps: Bayesialab (validation mode) > Analysis > Report > Target > Total Effects on Target > Quadrants.

It would be contrived to measure the correlation between the scores achieved by the students in their respective AI-ALS against their scores in the hardcopy paper-based post-test, because some students could have scored poorly in the AI-ALS as their poor understanding of certain math concepts might have been "surfaced" by the systems, but subsequently, they might have scored well in the paper-based post-test. Conversely, some students might have scored high in the AI-ALS because the questions were easy, but they might have scored low in the paper-based post-test. Hence, it absolutely does not mean that an AI-ALS would be ranked higher in the quadrant analysis chart if the students' scores within the AI-ALS are higher.

Each chart of the quadrant analysis generated by Bayesialab (see Figures 5 and 6) is divided into four quadrants. The variables' means (of each mathematics topic) are represented along the x-axis. The mean of the standardized total effect on the target (the paper-based post-test) is represented along the y-axis. Quadrant analysis example 1 (see Figure 5) utilized the file "data_five_classes_AI_ALS.csv". As a suggestion, the quadrants could be interpreted, as follows:

Top Right Quadrant (high volume, high impact on target node): This group contains the important variables with greater total effect on the target than the mean value. These AI-ALS are effective in contributing to the success of the students in the paper-based post-test. The AI-ALS supplied by Vendor 1, Vendor 2, Vendor 4, and Vendor 5 are in this category.

Top Left Quadrant (low volume, high impact on target node): Any AI-ALS in this category might be beneficial to the high-performing students, but not so beneficial to the mid- or low-performing students. There is no AI-ALS from any vendor in this quadrant.

Bottom Right Quadrant (high volume, low impact on target node): The AI-ALS from Vendor 3 is in this category, so educational stakeholders should consider conducting further investigation to find out why this AI-ALS could not contribute to beneficial results in the paper-based post-test for the students.

Bottom Left Quadrant (low volume, low impact on target node): Any AI-ALS in this category has relatively lower impact on the target node (the paper-based post-test). There is no AI-ALS from any vendor in this quadrant.

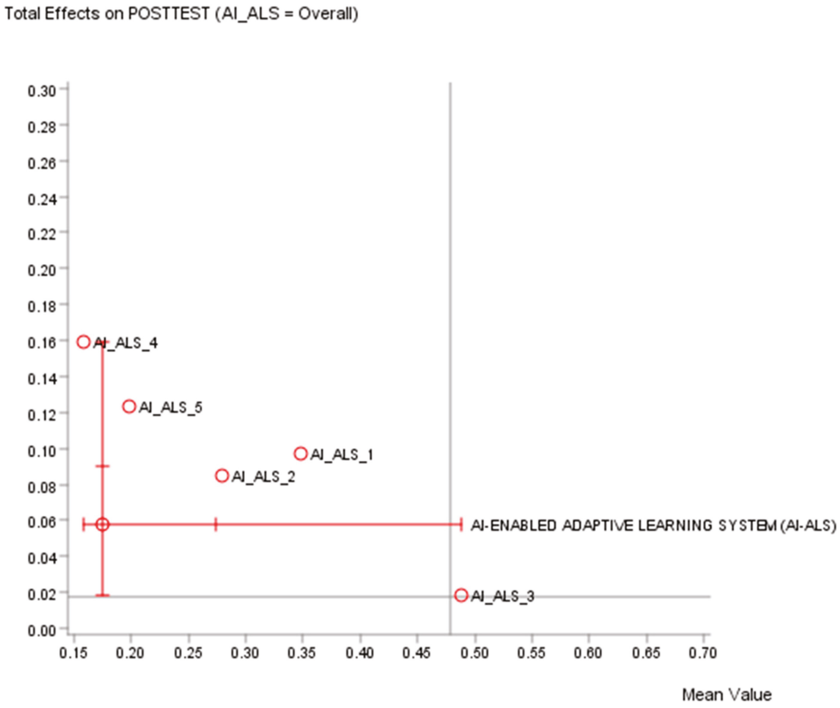


Figure 5. Comparison of Total Effects of the five different AI-ALS on the Post-test, which was machine-learned and generated by Bayesialab.

Quadrant analysis example 2 (see Figure 6) utilized the file "data_five_classes_AI_ALS.csv". As a suggestion, the quadrants could be interpreted, as follows:

Top Right Quadrant (high volume, high impact on target node): This quadrant contains the AI-ALS with greater total effect on the target than the mean value. Only the AI-ALS from Vendor 2 is in this quadrant. These noncognitive factors associated with this AI-ALS are important to the success of the students in the paper-based post-test, and the educational stakeholders should further explore how the noncognitive factors (e.g., motivation, stress management, psychological well-being, etc.) that are associated with the AI-ALS from Vendor 2 could be beneficial in helping the students to understand and learn the concepts well in these mathematics topics.

Top Left Quadrant (low volume, high impact on target node): Any AI-ALS in this category is associated with the noncognitive factors that might be beneficial for the high-performing students, but might not be so beneficial to the mid- or low-performing students. The AI-ALS supplied by Vendor 4 and Vendor 5 are in this quadrant.

Bottom Right Quadrant (high volume, low impact on target node): There is no AI-ALS from any vendor in this quadrant. If there is any AI-ALS in this category, educational stakeholders should consider conducting further investigation to find out why the noncognitive factors associated with this AI-ALS could not contribute to beneficial results in the paper-based post-test for the students.

Bottom Left Quadrant (low volume, low impact on target node): Any AI-ALS in this category has noncognitive factors that have relatively lower impact on the target node (the paper-based post-test). The AI-ALS from Vendor 1 and Vendor 3 are in this quadrant.

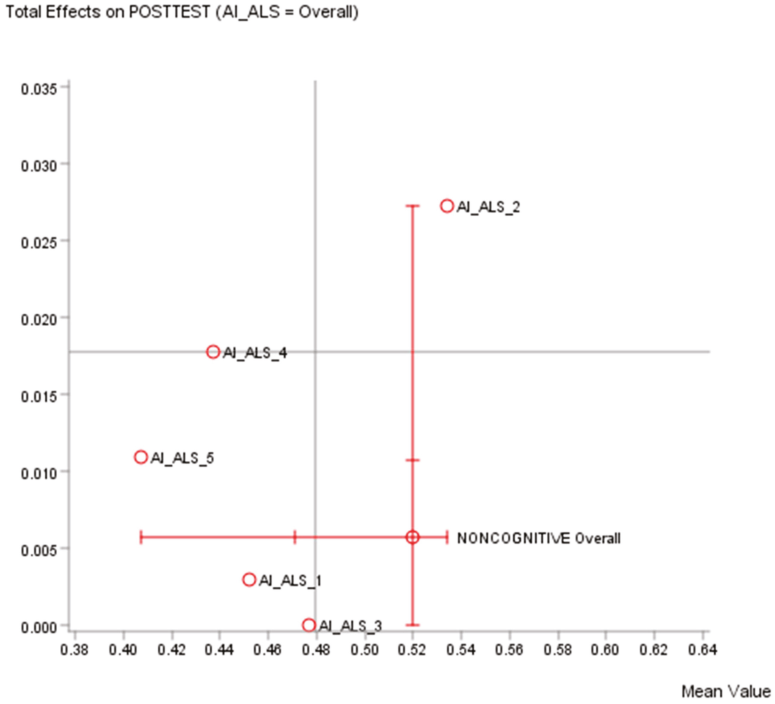


Figure 6. Comparison of Total Effects of the data in the Noncognitive node on the Post-test node, which was machine-learned from the data of the five different groups of students who had used five dissimilar AI-ALS.

4.7.3. Descriptive Analytics: Comparative Analysis of the Five AI-ALS

In this section, the performance results of the five classes of students who had used five dissimilar AI-ALS in five different schools will be presented.

Comparison between the AI-ALS from Vendor 1 and the Combined Average of the Five AI-ALS:

Using the file “data_ai_als_class_1.csv” via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 1 (see Figure 7):

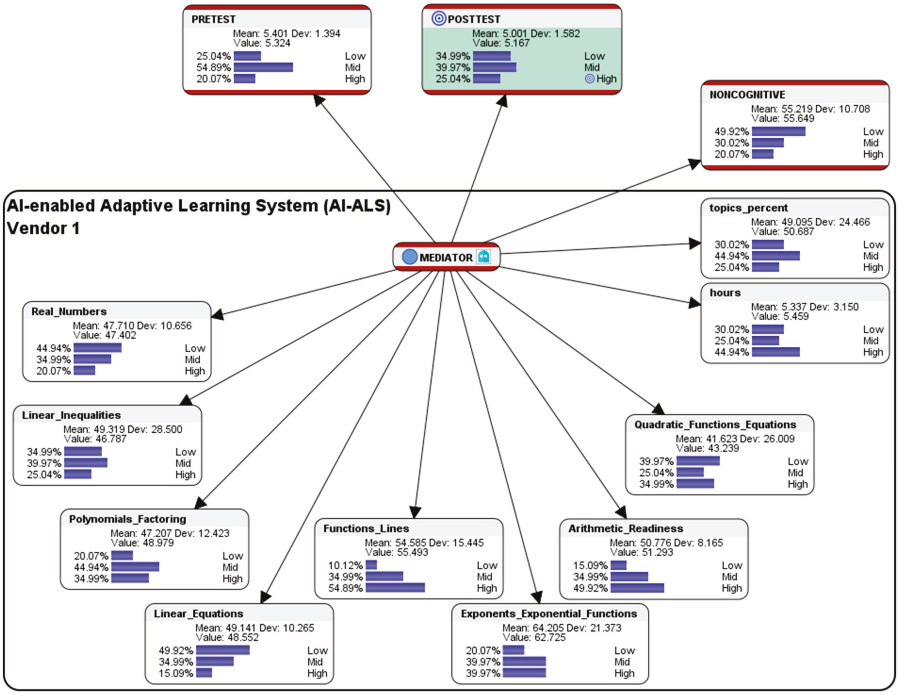


Figure 7. BN model of the students who had used the AI-ALS from Vendor 1 (N = 20 students).

In the paper-based Pre-test before the students used the AI-ALS from Vendor 1, 25.04% had scored at the Low-level (as compared to the combined average of 42% of the students who had scored at the Low-level), 54.89% had scored at the Mid-level (when compared to the combined average of 41% who had scored at the Mid-level), and 20.07% had scored at the High-level (as compared to the combined average of 17% scored at the High-level).

In the paper-based Post-test after the students had gone through the training within the AI-ALS from Vendor 1, 34.99% had scored at the Low-level (as compared to the combined average of 31% who had scored at the Low-level), 39.97% had scored at the Mid-level (when compared to the combined average of 47% who had scored at the Mid-level), and 25.04% had scored at the High-level (as compared to the combined average of 22% who had scored at the High-level). Overall, in terms of conventional gains by comparing the Pre-test vis-à-vis the Post-test, there was an unfavorable higher difference of 9.95% of the students who scored at the Low-level (from 25.04% in the Pre-test to 34.99% in the Post-test); there was a decline of 14.92% in the students who scored at the Mid-level (an decrease from 54.89% in the Pre-test to 39.97% in the Post-test); however, there was a favorable higher difference of 4.97% in the students who scored at the High-level (from 20.07% in the Pre-test to 25.04% in the Post-test).

In the aggregated Noncognitive factor, 49.92% of the students who had used the AI-ALS from Vendor 1 were at the so-called Low-level (a higher difference of 23.92% as compared to the combined average of 26% of the students who were at the Low-level), 30.02% were at the Mid-level (a lower difference of 12.98% as compared to the combined average of 43% of students who were at the Mid-level), and 20.07% were at the High-level (a lower difference of 10.93% when compared to the combined average of 31% of student who were at the High-level).

Within the AI-ALS from Vendor 1, in the topic of Real Numbers, 44.94% of the students scored at the Low-level (a higher difference of 16.94% as compared to the combined average of 28% of the students who scored at the Low-level), 34.99% of the students scored at the Mid-level (a lower difference of 10.01% as compared to the combined average of 45% of the students who scored at the Mid-level, and 20.07% of the students scored at the High-level (a lower difference of 6.93% compared to the combined average of 27% of the students who scored at the High-level.

In the topic of Linear Inequalities, 34.99% of the students scored at the Low-level (a higher difference of 1.99% compared to the combined average of 33% of the students who scored at the Low-level), 39.97% of the students scored at the Mid-level (a higher difference of 4.97% when compared to the combined average of 35% of the students who scored at the Mid-level), and 25.04% of the students scored at the High-level (a lower difference of 6.96% as compared to the combined average of 32% of the students who scored at the High-level.

In the topic of Polynomials and Factoring, 49.92% of the students scored at the Low-level (a higher difference of 35.92% when compared to the combined average of 14% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 12.01% as compared to the combined average of 47% of the students who scored at the Mid-level), and 15.09% scored at the High-level (a lower difference of 23.91% when compared to the combined average of 39% of the students who scored at the High-level).

In the topic of Linear Equations, 49.92% scored at the Low-level (a higher difference of 8.92% when compared to the combined average of 41% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 7.01% when compared to the combined average of 42% of the students who scored at the Mid-level), and 15.09% scored at the High-level (a lower difference of 1.91% when compared to the combined average of 17% scored at the High-level).

In the topic of Functions and Lines, 10.12% scored at the Low-level (a lower difference of 7.88% compared to the combined average of 18% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 6.01% as compared to the combined average of 41% of the students who scored at the Mid-level), and 54.89% who scored at the High-level (a higher difference of 13.89% as compared to the combined average of 41% of the students who scored at the High-level).

In the topic of Exponents and Exponential Functions, 20.07% scored at the Low-level (a higher difference of 16.93% when compared to the combined average of 37% of the students who scored at the Low-level), 39.97% scored at the Mid-level (a lower difference of 7.03% as compared to the combined average of 47% of the students who scored at the Mid-level), and 39.97% scored at the High-level (a higher difference of 23.97% when compared to the combined average of 16% of the students who scored at the High-level).

In the topic of Arithmetic Readiness, 15.09% scored at the Low-level (a higher difference of 3.09% compared to the combined average of 12% of the students who scored at the Low-level), 34.99% scored at the Mid-level (a lower difference of 20.01% compared to the combined average of 55% of the students who scored at the Mid-level), and 49.92% scored at the High-level (a higher difference of 16.92% s compared to the combined average of 33% scored at the High-level).

Regarding the topic of Quadratic Functions and Equations, 39.97% of the students scored at the Low-level (a higher difference of 16.97% as compared to the combined average of 23% of the students who scored at the Low-level), 25.04% scored at the Mid-level (a lower difference of 15.96% compared to the combined average of 41% scored at the Mid-level), and 34.99% scored at the High-level (a lower difference of 1.01% when compared to the combined average of 36% of the students who scored at the High-level).

Within the AI-ALS by Vendor 1, in the average number of hours spent by each student, 30.02% of the students were at the Low-level (a higher difference of 6.02% compared to the combined average of 24% of the students were at the Low-level), 25.04% were at the Mid-level (a lower difference of 8.96% as compared to the combined average of 34% of the students who were at the Mid-level), and 44.94% were at the High-level (a higher difference of 2.94% when compared to the combined average of 42% who were at the High-level).

In the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 1, 30.02% of the students were at the Low-level (a slightly lower difference of 0.98% compared to the combined average of 31% of the students who were at the Low-level), 44.94% were at the Mid-level (a higher difference of 4.94% compared to the combined average of 40% who were at the Mid-level), and 25.04% were at the High-level (a lower difference of 3.96% when compared to the combined average of 29% who were at the High-level).

Visualization of the Performance of the Students Who had Used Vendor 2's AI-ALS:

Using the file "data_ai_als_class_2.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 2 (see Figure 8):

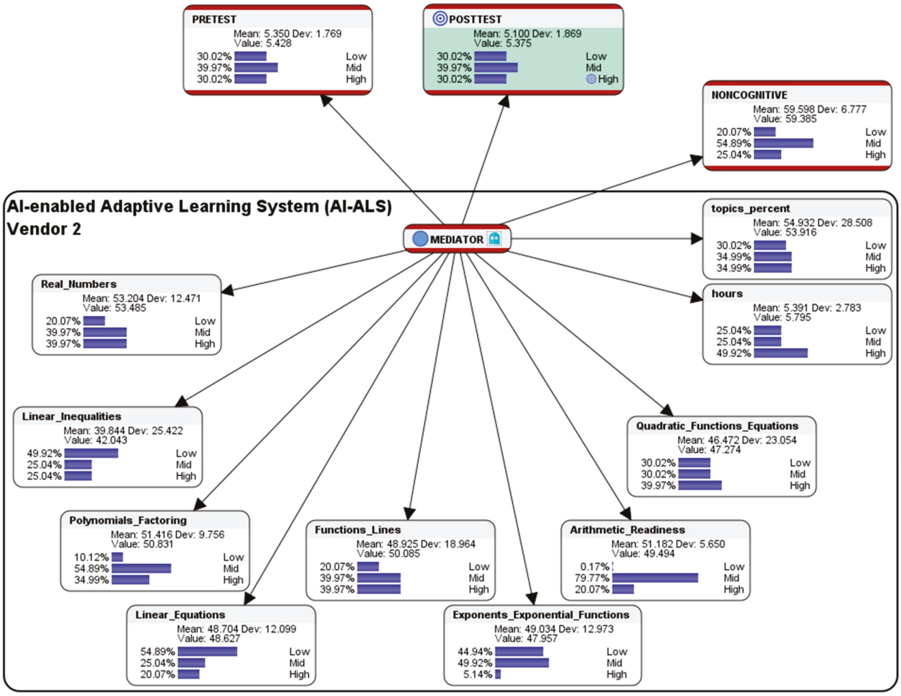


Figure 8. BN model of the students who had used the AI-ALS from Vendor 2 (N = 20 students).

Visualization of the Performance of the Students Who Had Used Vendor 3's AI-ALS:

Using the file "data_ai_als_class_3.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 3, (see Figure 9):

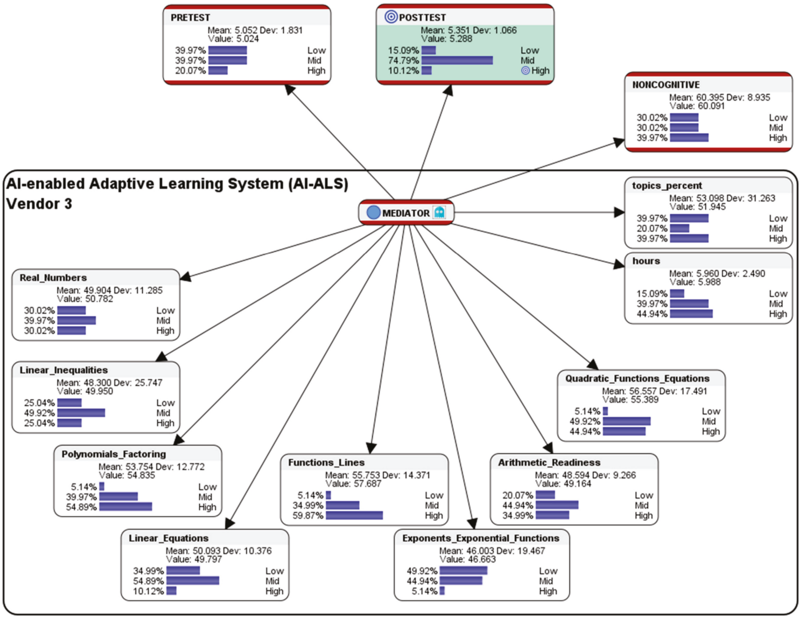


Figure 9. BN model of the students who had used the AI-ALS from Vendor 3 (N = 20 students).

Visualization of the Performance of the Students Who Had Used Vendor 4's AI-ALS:

Using the file "data_ai_als_class_4.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 4, (see Figure 10):

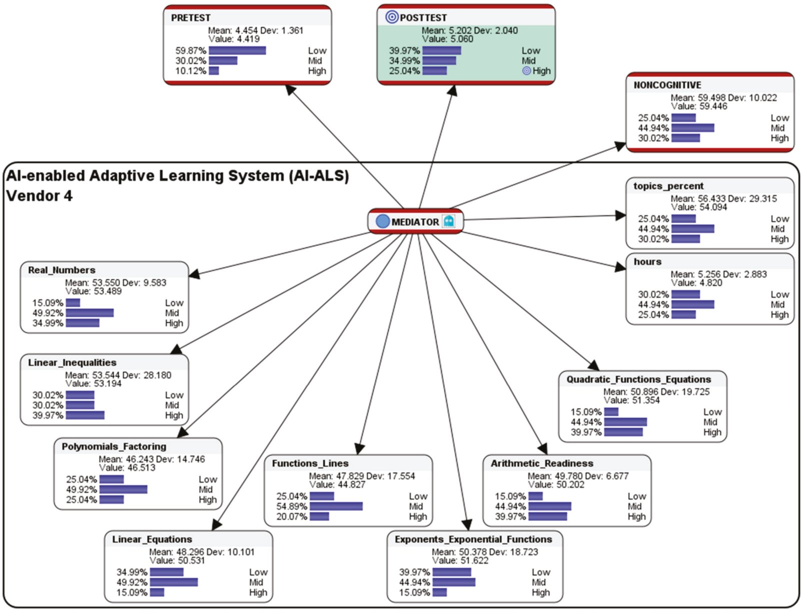


Figure 10. BN model of the students who had used the AI-ALS from Vendor 4 (N = 20 students).

Visualization of the Performance of the Students Who had Used Vendor 5's AI-ALS:

Using the file "data_ai_als_class_5.csv" via the Data Association tool in Bayesialab, the following score-clusters machine-learned by Bayesialab were observed from the dataset depicting the performances of the 20 students who had used the AI-ALS from Vendor 5 (see Figure 11):

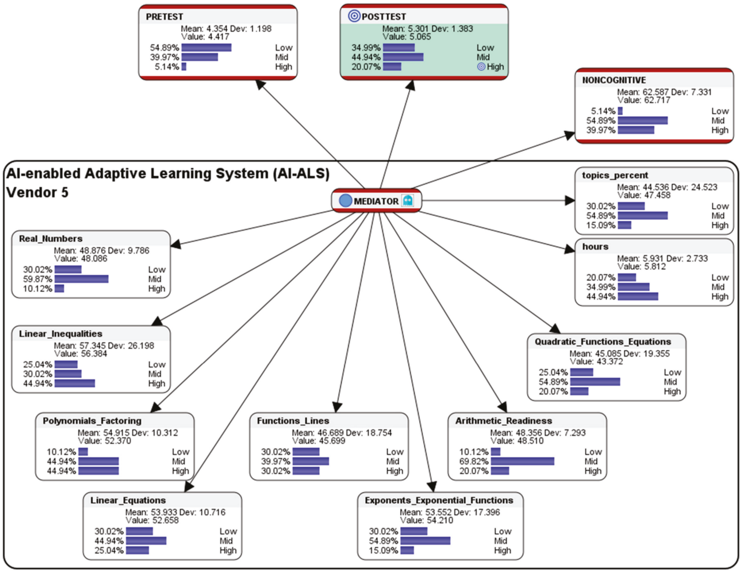


Figure 11. BN model of the students who had used the AI-ALS from Vendor 5 (N = 20 students).

4.7.4. Sensitivity Analysis of the Mathematics Topics that Contribute to the Performance of the Students who had Used the Five Dissimilar AI-ALS from the Five Vendors

Posterior Probability of the Post-test can be performed on the data from each school, while using tornado diagrams (see Figure 12). Sensitivity analysis can be activated in Bayesialab via these steps: *Bayesialab (validation mode) > Analysis > Visual > Sensitivity > Tornado diagrams on Total Effects*.

Each blue tornado chart of the total effects presents the performance (in the learning progress) of the students in each mathematics topic within the AI-ALS, in terms of the posterior probability of achieving high-level scores in the paper-based post-test. This implies that, in the AI-ALS proved by each vendor, the problem-solving practice that the students had in certain mathematics topics might have contributed to the high scores that were achieved by the students in the paper-based post-test. The longer blue bars represent higher sensitivity, in terms of how changes in the score of each mathematics topic (that is, their learning progress within each AI-ALS) could potentially affect the outcome in the paper-based post-test. Further coordination between the education stakeholders and the vendor of each respective AI-ALS should be carried out to understand how the teachers can focus on providing the students remediation of the more sensitive mathematics topics (represented with longer blue bars), as they seem to be important in affecting the performance of their students who could score high marks in the paper-based post-test.

Each red tornado chart of the total effects presents the performance of the students in each mathematics topic within the AI-ALS, in terms of the posterior probability of achieving low-level scores in the paper-based post-test. This implies that, in the AI-ALS proved by the vendor, the problem-solving practice that the students had in the mathematics topics might have contributed to the high scores that were achieved by the students in the paper-based post-test. The longer red bars represent higher sensitivity, in terms of how changes in the score of each mathematics topic (that is, their learning progress within each AI-ALS) could potentially affect the outcome in the paper-based post-test. Further coordination via discussions between the education stakeholders and each respective vendor of the AI-ALS should be carried out to understand how the teachers can focus on providing the students remediation of the more sensitive mathematics topics (represented with longer red bars), as they seem to be affecting the performance of their students who could only score low marks in the paper-based post-test.

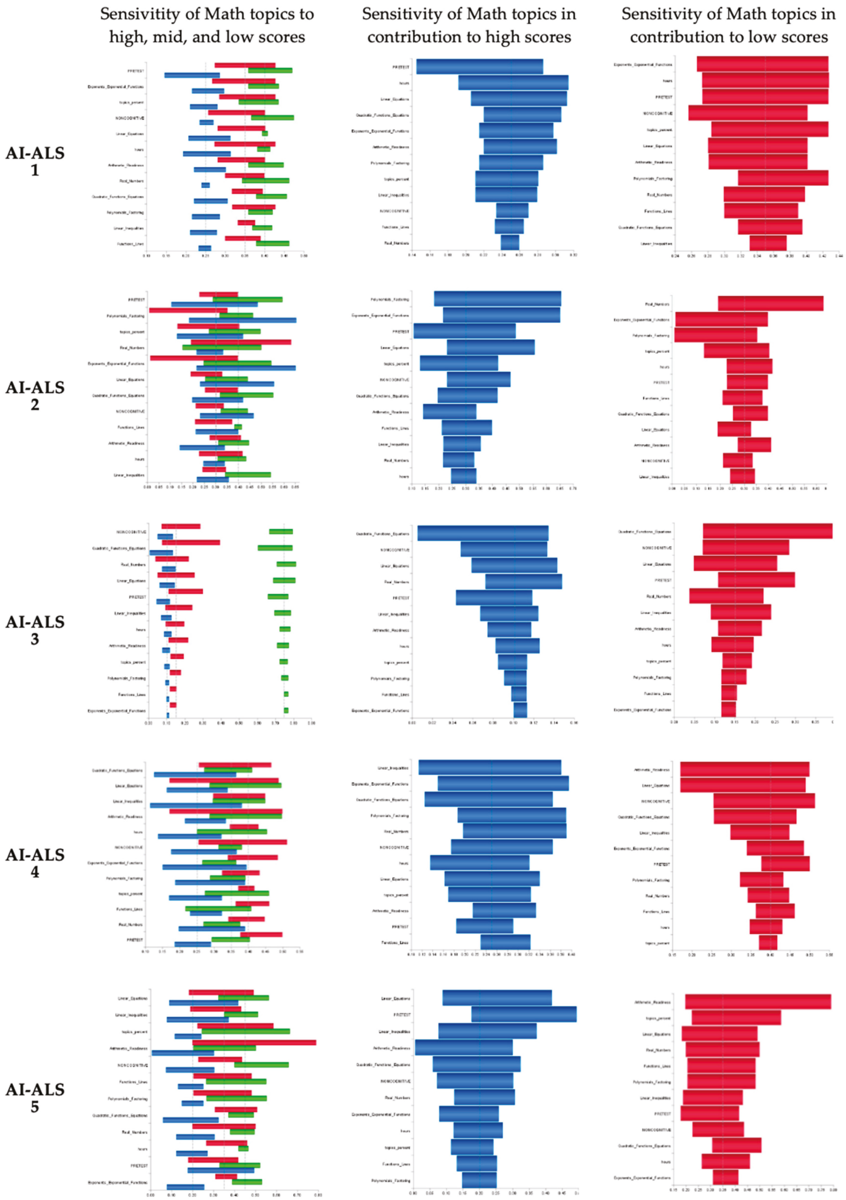
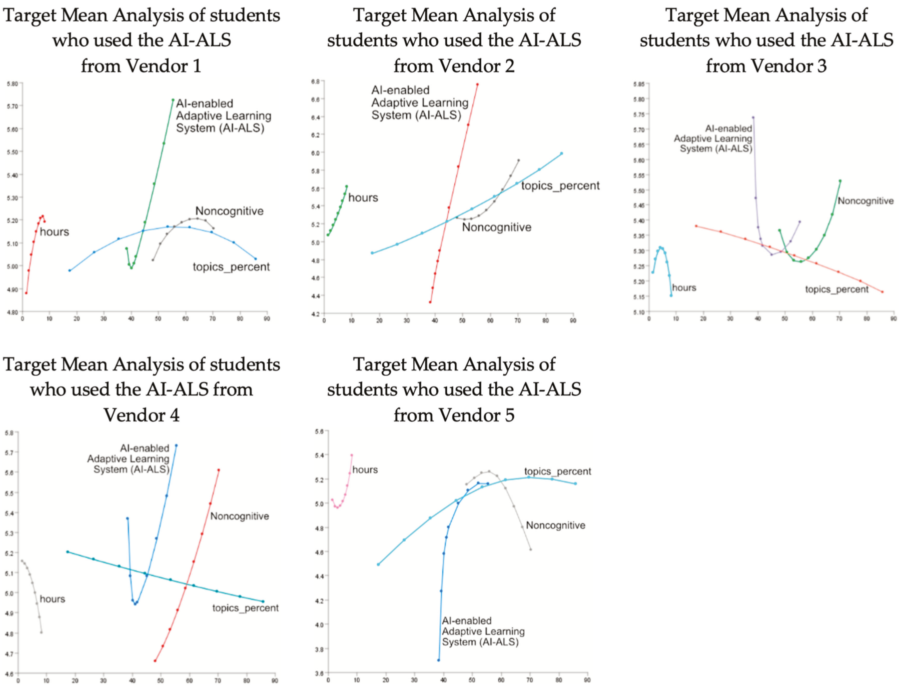


Figure 12. Visualizations of the sensitivity analysis data of the five groups of students in their respective AI-ALS, regarding how their learning progress of the mathematics topics within each AI-ALS could potentially affect their outcomes in the paper-based post-test.

4.7.5. Descriptive Analytics: Oversight Using Curves Analysis of the AI-ALS from the Five Vendors

Another way to visualize the influence of the students' mastery of the various mathematics topics on their paper-based post-test can be accomplished by using this tool in Baysialab via these steps on the menubar: Bayesialab (validation mode) > Analysis > Visual > Target > Target's Posterior > Curves > Total Effects.

As observed in Figure 13, the plots of the data reveal that the relationships between the total effects and the various factors on the target node (that is, the paper-based post-test) could be linear or curvilinear. The curvilinear lines suggest that there might be "peaks" or "valleys" in some of the relationships between the input variables (e.g., the number of hours spent using the AI-ALS, or the quality of the noncognitive factors, or the scores achieved by the students within each AI-ALS, or the percentage of mathematics topics mastered within the AI-ALS) and their respective educational outcomes in the paper-based post-test. With these curves analysis charts, further discussions could be initiated amongst the policy makers, technology vendors, teachers, parents, and students to help improve the learning experiences of the students.



Target Mean Analysis of students who used the AI-ALS from Vendor 1

Target Mean Analysis of students who used the AI-ALS from Vendor 2

Target Mean Analysis of students who used the AI-ALS from Vendor 3

Target Mean Analysis of students who used the AI-ALS from Vendor 4

Target Mean Analysis of students who used the AI-ALS from Vendor 5

Figure 13. Target Mean Analysis of five different groups of students, each of which had used an AI-ALS from a different vendor.

4.7.6. Descriptive Analytics: Pearson Correlation Analysis

Descriptive analytics can also be performed using the Pearson correlation analysis tool in Bayesialab. It can be used for the corroboration of the relationship analyses between the students' learning performances in the AI-ALS and their corresponding performances in the paper-based post-test. The visualizations of the Pearson correlations can be presented, so that it is easier to see the positive correlations highlighted in blue, and the negative correlations faded out in red (see Figure 14). This tool can be activated in Bayesialab via these steps on the menubar: Analysis > Visual > Overall > Arc > Pearson Correlation.

One suggestion for interpretation of the negative Pearson correlations could be that the red lines and nodes might represent the regions where the weaknesses of the students were "surfaced" or educed (drawn out) by the AI-ALS. It might not necessarily be an undesirable situation, provided that the teacher could provide remediation to the students so that the gaps that the AI-ALS could not bridge for the students (e.g., if the AI-ALS could not read the students' workings to pin-point where the mathematical calculation mistakes were for the students) were addressed.

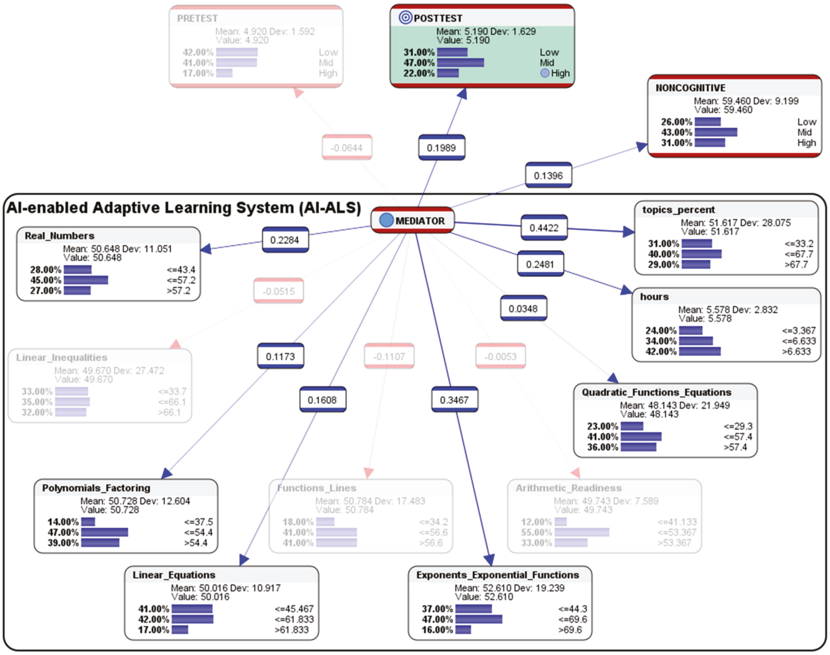


Figure 14. Pearson correlations between the students' learning progress of the mathematics topics within the AI-ALS and their corresponding performances in the paper-based post-test.

4.7.7. Descriptive Analytics: Oversight of the Gains in the Different Groups of Students

No gain in performance (scores in the post-test vis-à-vis the pre-test) was observed for the students who had used AI-ALS from Vendor 2, and negative gain (the scores in the post-test were lower than those in the pre-test) was observed for the students who had used the AI-ALS from Vendor 3, as observed in Table 1 and Figure 15. However, it might not be the fault of the AI-ALS that those students underperformed. Further qualitative interviews with the students might reveal the possible reasons for these preliminary observations.

Table 1. Comparisons between scores within the five AI-ALS and the paper-based post-tests.

AI-ALS Vendor	AI-ALS Low-Level Score (% of Students)	AI-ALS High-Level Score (% of Students)	Post-Pre Test High-Level Score Gain (% of Students)
1	35.00	30.10	4.97
2	50.10	29.89	0.00
3	25.04	44.24	−9.95
4	35.06	30.05	14.92
5	29.63	45.32	14.93

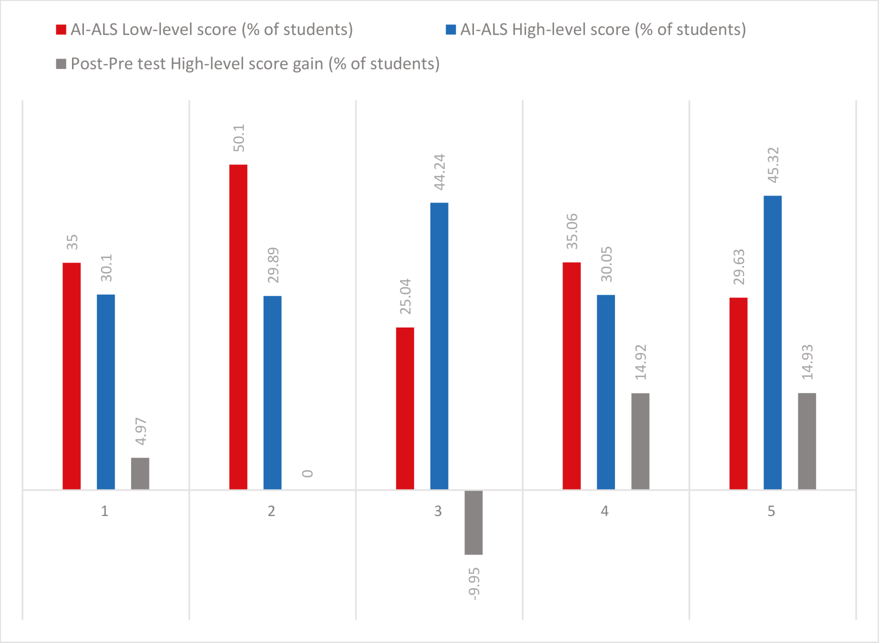


Figure 15. Histograms depicting the performance of each class of students: the low-level scores within each AI-ALS are presented in red; the high-level scores within each AI-ALS are presented in blue; their corresponding high-level score gains in the paper-based post-test are represented in gray.

There seemed to be no clear pattern of correlation between the difficulty of scoring high-level scores or low-level scores within each AI-ALS and the gains in the high-level scores in the paper-based post-test, contrary to what was initially hypothesized by the researcher in Section 2.2. In other words, making it easy (or even difficult) for the students to score at the high-level might not necessarily result in corresponding high-level gains in the paper-based post-test, probably because of the uniqueness of each AI-ALS and each class of students.

However, although direct comparisons between the five AI-ALS might seem challenging, it would still be possible to predict how the performance of each group of students within their respective AI-ALS could be optimized to achieve high scores in the paper-based post-test. To demonstrate that, "what-if?" predictive analytics would be utilized in the subsequent section.

5. "What-If?" Predictive Analytics

In this section, the following predictive analytics reports will be presented unabridged, in order to delineate how human-centric reasoning could be applied to interpret the counterfactual results that were generated by the AI-based BN model. For better readability, the first letter of the names of the BN nodes and entities would be presented in the upper case.

5.1. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 1

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 1, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 1 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

To predict the conditions that would enable 100% of the students in Class 1, who had used Vendor 1's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of score-clusters were observed (see Figure 16):

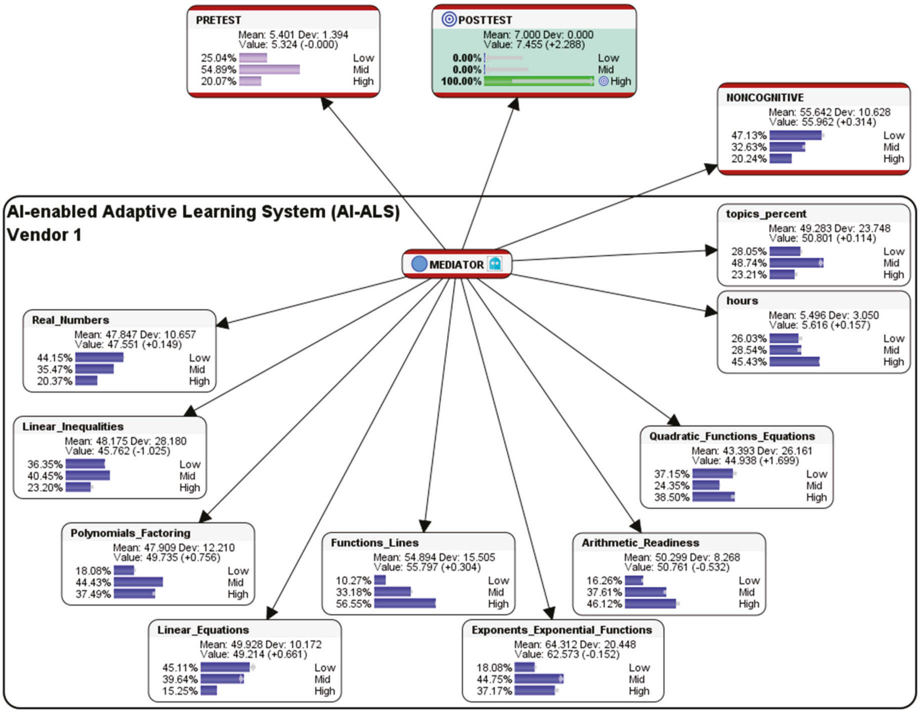


Figure 16. Simulation of counterfactual results for 100% of the students who had used Vendor 1's AI-ALS to score at the high-level in the post-test.

Within the AI-ALS from Vendor 1, in the aggregated Noncognitive factor, ideally 47.13% of the students who had used the AI-ALS from Vendor 1 should be at the so-called Low-level (a lower difference of 2.79% when compared to the original 49.92% of the students who were at the Low-level); 32.63% should be at the Mid-level (a higher difference of 2.61% compared to the original 30.02% of students who were at the Mid-level); and 20.64% should be at the High-level (an almost negligible higher difference of 0.57% as compared to the original 20.07% of students who were at the High-level).

Within the AI-ALS from Vendor 1, in the topic of Real Numbers, ideally 44.15% of the students should score at the Low-level (a slightly lower difference of 0.79% compared to the original 44.94% of the students who scored at the Low-level), 35.47% of the students should score at the Mid-level (a slightly higher difference of 0.48% as compared to the original 34.99% of the students who scored at the Mid-level), and 20.37% of the students should score at the High-level (a slightly higher difference of 0.3% when compared to the original 20.07% of the students who scored at the High-level. The simulated results for the topic of Real Numbers suggest that Vendor 1's AI-ALS was already performing close to optimum in terms of contributing the students scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Linear Inequalities, ideally 36.35% of the students should score at the Low-level (a higher difference of 1.36% as compared to the original 34.99% of the students who scored at the Low-level); 40.45% of the students should score at the Mid-level (an almost negligible higher difference of 0.48% when compared to the original 39.97% of the students who scored at the Mid-level); and, 23.20% of the students should score at the High-level (a slightly lower difference of 1.84% as compared to the original 25.04% of the students who scored at the High-level. The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Polynomials and Factoring, ideally 18.08% of the students should score at the Low-level (a substantially lower difference of 16.91% as compared to the original 49.92% of the students who scored at the Low-level); 44.43% should score at the Mid-level (a higher difference of 9.44% when compared to the original 34.99% of the students who scored at the Mid-level); and, 37.49% should score at the High-level (a substantially higher difference of 22.40% as compared to the original 15.09% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Linear Equations, ideally 45.11% of the students should score at the Low-level (a lower difference of 4.81% when compared to the original 49.92% of the students who scored at the Low-level); 39.64% should score at the Mid-level (a lower difference of 4.65% when compared to the original 34.99% of the students who scored at the Mid-level); and, 15.25% should score at the High-level (an almost negligible higher difference of 0.16% when compared to the original 15.09% that scored at the High-level). The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Functions and Lines, ideally 10.27% of the students should score at the Low-level (an almost negligible higher difference of 0.15% as compared to the original 10.12% of the students who scored at the Low-level); 33.18% should score at the Mid-level (a lower difference of 1.81% when compared to the original 34.99% of the students who scored at the Mid-level); and, 56.55% should score at the High-level (a higher difference of 1.66% compared to the original 54.89% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Exponents and Exponential Functions, ideally 18.08% should score at the Low-level (a lower difference of 1.99% as compared to the original 20.07% of the students who scored at the Low-level); 44.75% should score at the Mid-level (a higher difference of 4.78% when compared to the original 39.97% of the students who scored at the Mid-level); and, 37.17% should score at the High-level (a lower difference of 2.8% compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Arithmetic Readiness, ideally 16.26% should score at the Low-level (a slightly higher difference of 1.17% as compared to the original 15.09% of the students who scored at the Low-level); 37.61% should score at the Mid-level (a higher difference of 2.62% when compared to the original 34.99% of the students who scored at the Mid-level); and, 46.12% should score at the High-level (a lower difference of 3.8% compared to the original 49.92% scored at the High-level). The simulated results suggest that, if Vendor 1′s AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the topic of Quadratic Functions and Equations, ideally 37.15% of the students should score at the Low-level (a lower difference of 2.82% as compared to the original 39.97% of the students who scored at the Low-level); 24.35% should score at the Mid-level (an almost negligible lower difference of 0.69% when compared to the original 25.04% who scored at the Mid-level); and, 38.50% should score at the High-level (a higher difference of 3.51% as compared to the original 34.99% of the students who scored at the High-level). The simulated results suggest that if Vendor 1's AI-ALS could ideally make it slightly easier for students in Class 1 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 1, in the average number of hours spent by each student, ideally 26.03% of the students should be at the Low-level (a lower difference of 3.99% as compared to the original 30.02% of the students who were at the Low-level); 28.54% should be at the Mid-level (a higher difference of 3.5% when compared to the original 25.04% of the students who were at the Mid-level); and 45.43% should be at the High-level (an almost negligible higher difference of 0.49% as compared to the original 44.94% who were at the High-level). The simulated results suggest that more time spent using the AI-ALS might contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 1, in the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 1, ideally 28.05% of the students should be at the Low-level (a slightly lower difference of 1.97% as compared to the original 30.02% of the students who were at the Low-level); 48.74% should be at the Mid-level (a higher difference of 3.8% compared to the original 44.94% who were at the Mid-level); and, 23.21% should be at the High-level (a lower difference of 1.83% when compared to the original 25.04% who were at the High-level). The simulated results suggest that Vendor 1's AI-ALS was effective in providing adaptive learning to the students and was contributing well to their probability of scoring high marks in the paper-based Post-test.

5.2. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 2

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 2, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 2 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

To predict the conditions that would enable 100% of the students in Class 2 who had used Vendor 2's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of the score-clusters were observed (see Figure 17):

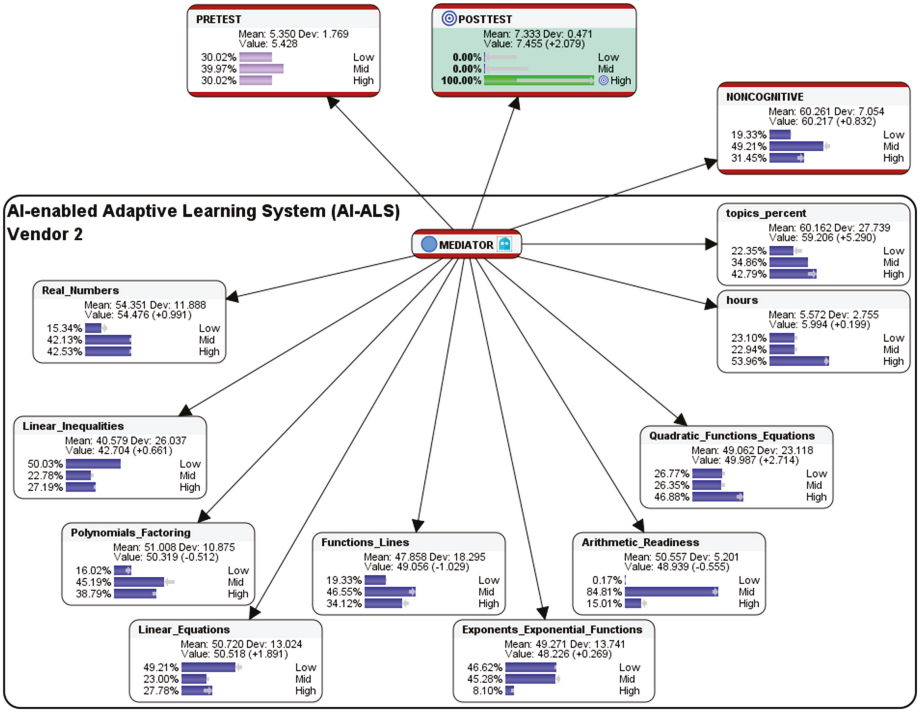


Figure 17. Simulation of counterfactual results for 100% of the students who had used Vendor 2's AI-ALS to score at the high-level in the post-test.

Within the AI-ALS from Vendor 2, in the aggregated Noncognitive factor, ideally 19.33% of the students who had used the AI-ALS from Vendor 2 should be at the so-called Low-level (an almost negligible lower difference of 0.74% as compared to the original 20.07% of the students who were at the Low-level); 49.21% should be at the Mid-level (a higher difference of 5.68% when compared to the original 54.89% of students who were at the Mid-level); and, 31.45% should be at the High-level (a higher difference of 6.41% as compared to the original 25.04% of students who were at the High-level). The counterfactual results suggest that, if the mid-level and high-level of noncognitive attributes (e.g., emotional intelligence to manage stress, interest in learning mathematics, motivation, level of engagement, etc.) could be increased, it might contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Real Numbers, ideally 15.34% of the students should score at the Low-level (a slightly lower difference of 4.73% as compared to the original 20.07% of the students who scored at the Low-level); 42.13% of the students should score at the Mid-level (a slightly higher difference of 2.16% when compared to the original 39.97% of the students who scored at the Mid-level); and, 42.53% of the students should score at the High-level (a slightly higher difference of 2.56% as compared to the original 39.97% of the students who scored at the High-level. The simulated counterfactual results for the topic of Real Numbers suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Linear Inequalities, ideally 50.03% of the students should score at the Low-level (an almost negligible higher difference of 0.11% as compared to the original 49.92% of the students who scored at the Low-level); 22.78% of the students should score at the Mid-level (a slightly lower difference of 2.26% when compared to the original 25.04% of the students who scored at the Mid-level); and, 27.19% of the students should score at the High-level (a slightly higher difference of 2.15% as compared to the original 25.04% of the students who scored at the High-level. The simulated counterfactual results for the topic suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Polynomials and Factoring, ideally 16.02% of the students should score at the Low-level (a lower difference of 5.90% as compared to the original 10.12% of the students who scored at the Low-level); 45.19% should score at the Mid-level (a substantially lower difference of 9.70% when compared to the original 54.89% of the students who scored at the Mid-level); and, 38.79% should score at the High-level (a slightly higher difference of 3.80% compared to the original 34.99% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Linear Equations, ideally 49.21% of the students should score at the Low-level (a lower difference of 5.68% when compared to the original 54.89% of the students who scored at the Low-level); 23.00% should score at the Mid-level (a lower difference of 2.04% when compared to the original 25.04% of the students who scored at the Mid-level); and 27.78% should score at the High-level (a higher difference of 7.71% when compared to the original 20.07% who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Functions and Lines, ideally 19.33% of the students should score at the Low-level (an almost negligible lower difference of 0.74% as compared to the original 20.07% of the students who scored at the Low-level); 46.55% should score at the Mid-level (a higher difference of 6.58% when compared to the original 39.97% of the students who scored at the Mid-level); and, 34.12% should score at the High-level (a lower difference of 5.85% compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that if Vendor 2's AI-ALS could ideally make it more difficult for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Exponents and Exponential Functions, ideally 46.62% should score at the Low-level (a slightly higher difference of 1.68% when compared to the original 44.94% of the students who scored at the Low-level); 45.28% should score at the Mid-level (a lower difference of 4.64% as compared to the original 49.92% of the students who scored at the Mid-level); and, 8.10% should score at the High-level (a lower difference of 2.96% compared to the original 5.14% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly easier for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Arithmetic Readiness, ideally 0.17% should score at the Low-level (a difference of 0.00% as compared to the original 0.17% of the students who scored at the Low-level); 84.81% should score at the Mid-level (a higher difference of 5.04% compared to the original 79.77% of the students who scored at the Mid-level); and, 15.01% should score at the High-level (a lower difference of 5.06% compared to the original 20.07% who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the topic of Quadratic Functions and Equations, ideally 26.77% of the students should score at the Low-level (a lower difference of 3.25% when compared to the original 30.02% of the students who scored at the Low-level); 26.35% should score at the Mid-level (a lower difference of 3.67% as compared to the original 30.02% who scored at the Mid-level); and, 46.88% should score at the High-level (a higher difference of 6.91% when compared to the original 39.97% of the students who scored at the High-level). The simulated results suggest that, if Vendor 2's AI-ALS could ideally make it easier for students in Class 2 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 2, in the average number of hours spent by each student, ideally 23.10% of the students should be at the Low-level (a lower difference of 1.94% as compared to the original 25.04% of the students who were at the Low-level); 22.94% should be at the Mid-level (a slightly lower difference of 2.10% when compared to the original 25.04% of the students who were at the Mid-level), and, 53.96% should be at the High-level (a slightly higher difference of 4.04% as compared to the original 49.92% who were at the High-level). The simulated results suggest that if the students could spend more time learning mathematics within Vendor 2's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test.

Within the AI-ALS from Vendor 2, in the percentage of the total number of topics that were mastered by the students, ideally 22.35% of the students should be at the Low-level (a slightly higher difference of 7.67% as compared to the original 30.02% of the students who were at the Low-level); 34.86% should be at the Mid-level (an almost negligible lower difference of 0.13% compared to the original 34.99% who were at the Mid-level); and, 42.79% should be at the High-level (a higher difference of 7.8% compared to the original 34.99% who were at the High-level). The simulated results suggest that if the students could master a higher percentage of topics within Vendor 2's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test.

5.3. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 3

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 3, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions that are needed in the AI-ALS from Vendor 3 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

Here is an opportunity that the following analysis can be used as a starting point for discussions to foster strategic coordination between the educational stakeholders and Vendor 3 which provided the AI-ALS. As previously observed in Table 1 and Figure 15, there was a decrease in the number of students who scored at the High-level of the marks in the paper-based post-test. Realistically, since the algorithm with which the AI-ALS from Vendor 3 interacts with the students cannot be changed much, if at all, the mathematics teacher would have to provide remediation for the students. The AI-ALS from Vendor 3 might not be a good choice in the selection for in-service deployment from the perspective of the policy makers and educational stakeholders, as it might be realistically impractical to ask Vendor 3 to change their proprietary algorithm to suit the students of Class 3. However, the simulated counterfactual results (see Figure 18) could still be used as a guide for remediation by the teacher to "level-up" the students in the mathematics topics that they might be weaker in.

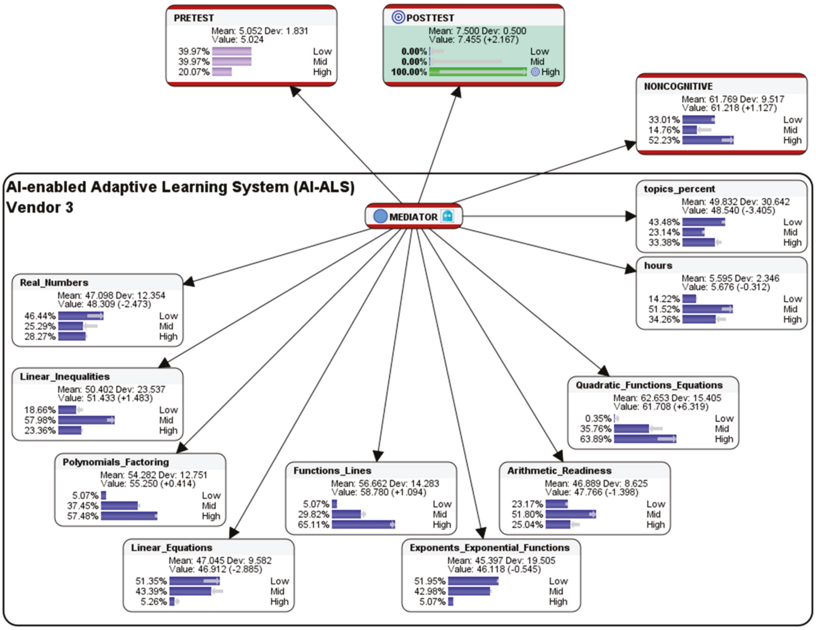


Figure 18. Simulation of counterfactual results for 100% of the students who had used Vendor 3's AI-ALS to score at the high-level in the post-test.

To predict the conditions that would enable 100% of the students in Class 3 who had used Vendor 3's AI-ALS to score at the High-level in the paper-based Post-test, hard evidence was set on it (by double-clicking on the High-level histogram bar in Bayesialab). The following counterfactually simulated results of score-clusters were observed (see Figure 18):

In the aggregated Noncognitive factor, ideally 33.01% of the students who had used the AI-ALS from Vendor 3 should be at the so-called Low-level (a slightly higher difference of 2.99% compared to the original 30.02% of the students who were at the Low-level); 14.76% should be at the Mid-level (a substantially lower difference of 15.26% as compared to the original 30.02% of students who were at the Mid-level); and, 50.23% should be at the High-level (a substantially higher difference of 10.26% as compared to the original 39.97% of students who were at the High-level). The results suggest that noncognitive factors of the students such as motivation, interest, attitude towards mathematics, etc. might need to be improved.

Within the AI-ALS from Vendor 3, in the topic of Real Numbers, ideally 46.44% of the students should score at the Low-level (a substantially higher difference of 16.42% when compared to the original 30.02% of the students who scored at the Low-level), 25.29% of the students should score at the Mid-level (a substantially lower difference of 14.68% as compared to the original 39.97% of the students who scored at the Mid-level), and 28.27% of the students should score at the High-level (a slightly lower difference of 1.75% when compared to the original 30.02% of the students who scored at the High-level. The simulated results suggest that, if Vendor 3's AI-ALS could ideally make it slightly more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Linear Inequalities, ideally 18.66% of the students should score at the Low-level (a lower difference of 6.38% when compared to the original 25.04% of the students who scored at the Low-level); 57.98% of the students should score at the Mid-level (a higher difference of 8.06% as compared to the original 49.92% of the students who scored at the Mid-level); and, 23.36% of the students should score at the High-level (a slightly lower difference of

9.74% as compared to the original 25.04% of the students who scored at the High-level. The simulated results suggest that, ideally, if Vendor 3's AI-ALS could make it slightly more difficult for students to score at the High-level, but yet, not so difficult that students find it too challenging to score at the Mid-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Polynomials and Factoring, ideally 5.07% of the students should score at the Low-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the Low-level); 37.45% should score at the Mid-level (an almost negligible lower difference of 0.04% as compared to the original 39.97% of the students who scored at the Mid-level); and, 57.48% should score at the High-level (a slightly higher difference of 2.59% when compared to the original 54.89% of the students who scored at the High-level). The simulated results suggest that Vendor 3's AI-ALS might already be close to optimally adapting to the students in Class 3 in training them to score at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Linear Equations, ideally 51.35% of the students should score at the Low-level (a substantially higher difference of 16.36% when compared to the original 34.99% of the students who scored at the Low-level); 43.39% should score at the Mid-level (a substantially lower difference of 11.5% when compared to the original 54.89% of the students who scored at the Mid-level); and, 5.26% should score at the High-level (a lower difference of 4.86% when compared to the original 10.12% scored at the High-level). The simulated results suggest that, if Vendor 3's AI-ALS could ideally make it much more difficult for students to score at the High-level and at the Mid-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Functions and Lines, ideally 5.07% of the students should score at the Low-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the Low-level); 29.82% should score at the Mid-level (a lower difference of 5.17% as compared to the original 34.99% of the students who scored at the Mid-level); and, 65.11% should score at the High-level (a higher difference of 5.24% compared to the original 59.87% of the students who scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it slightly easier for students in Class 3 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Exponents and Exponential Functions, ideally 51.95% should score at the Low-level (a lower difference of 2.03% as compared to the original 49.92% of the students who scored at the Low-level); 42.98% should score at the Mid-level (a lower difference of 6.94% when compared to the original 49.92% of the students who scored at the Mid-level); and, 5.07% should score at the High-level (an almost negligible lower difference of 0.07% when compared to the original 5.14% of the students who scored at the High-level). The simulated results suggest that Vendor 3's AI-ALS might already be close to optimally adapting to the students in Class 3 in training them to score at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Arithmetic Readiness, ideally 23.17% should score at the Low-level (a slightly higher difference of 3.1% as compared to the original 20.07% of the students who scored at the Low-level); 51.80% should score at the Mid-level (a higher difference of 6.86% when compared to the original 44.94% of the students who scored at the Mid-level); and, 25.04% should score at the High-level (a lower difference of 9.95% as compared to the original 34.99% scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it much more difficult for students to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS from Vendor 3, in the topic of Quadratic Functions and Equations, ideally 0.35% of the students should score at the Low-level (a lower difference of 4.79% as compared to the original 5.14% of the students who scored at the Low-level); 35.76% should score at the Mid-level (a substantially lower difference of 14.16% when compared to the original 49.92% who scored at the Mid-level); and, 63.89% should score at the High-level (a substantially higher difference of 18.95% as compared to the original 44.94% of the students who scored at the High-level). The simulated results suggest that if Vendor 3's AI-ALS could ideally make it much easier for students in Class 3 to score at the High-level, it might contribute to their probability of scoring at the High-level for this topic in the paper-based Post-test.

Within the AI-ALS by Vendor 3, in the average number of hours were spent by each student, 14.22% of the students should be at the Low-level (an almost negligible lower difference of 0.87% compared to the original 15.09% of the students who were at the Low-level), 51.52% should be at the Mid-level (a substantially higher difference of 11.55% as compared to the original 39.97% of the students who were at the Mid-level), and 34.26% should be at the High-level (a substantially lower difference of 10.68% compared to the original 44.94% who were at the High-level). The simulated results suggest that if students spend less time within Vendor 3's AI-ALS, it could contribute to their probability of scoring at the High-level in the paper-based Post-test. Perhaps, one way of interpreting this could be: spending less time within the Vendor 3's AI-ALS could help prevent diminishing marginal returns, as the students would not have to suffer from undue fatigue or stress.

In the percentage of the total number of topics that were mastered by the students in the AI-ALS by Vendor 3, ideally 43.48% of the students should be at the Low-level (a slightly lower difference of 3.51% as compared to the original 39.97% of the students who were at the Low-level); 23.14% should be at the Mid-level (a slightly higher difference of 3.07% when compared to the original 20.07% who were at the Mid-level); and 33.38% should be at the High-level (a lower difference of 6.59% as compared to the original 39.97% who were at the High-level). The simulated results suggest that mastering the topics at a slower pace within Vendor 3's AI-ALS could contribute to their probability of scoring at the High-level in the paper-based Post-test. At first glance, this might seem counterintuitive. However, one way of interpreting this might be: a slower pace of mastering the mathematics topics could be more beneficial, as it could potentially contribute to a deeper level of understanding of the subject matter by the students.

5.4. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 4

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 4, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions needed in the AI-ALS from Vendor 4 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the paper-based Post-test?

The following simulated counterfactual results for the conditions that would optimize the performance of students who had used the AI-ALS from Vendor 4 (see Figure 19) and Vendor 5 (see Figure 20) would only be presented in summarized graphical form due to space constraints for publication, since they also had positive gains in the High-level marks in the Post-test (as presented earlier in Table 1 and Figure 15), and they could be considered to be similar to the case in which the students had used the AI-ALS from Vendor 1 (see Section 5.1).

Overall, within the AI-ALS from Vendor 4, the simulated counterfactual results suggest that, in order to train them in score at the High-level in the paper-based Post-test, the finer details of the predictions that recommend whether it should be made easier or more difficult in the various mathematics topics could be perused in Figure 19.

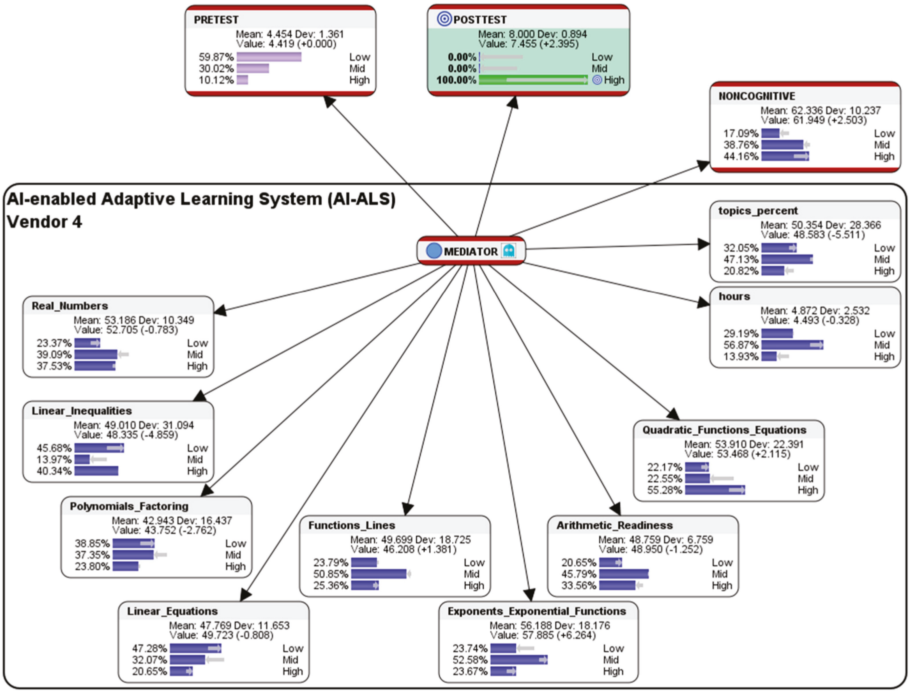


Figure 19. Simulation of counterfactual results for 100% of the students who had used Vendor 4's AI-ALS to score at the high-level in the post-test. Grey arrows recommended whether there should be an increase (pointing to the right), or a decrease (pointing to the left) in each respective mathematics topic's score-clusters for Low-, Mid-, and High-level.

5.5. Simulation of Hypothetical Scenario for Students Who had Used the AI-ALS from Vendor 5

This section presents a sample performance prediction report that could be shared with the educational stakeholders in School 5, so that they could consider having further discussions with their AI-ALS provider to fine-tune the system, e.g., by adjusting the level of difficulty of the questions that are being offered to their students to better correspond to their learning capabilities.

Hypothetical question: what are the conditions that are needed in the AI-ALS from Vendor 5 and in the noncognitive parameter *if* we wish that 100% of the students could score at the High-level in the Post-test?

Overall, within the AI-ALS from Vendor 5, the simulated counterfactual results suggest that, in order to train them in score at the High-level in the paper-based Post-test, the finer details of the predictions that recommend whether it should be made easier or more difficult for the various mathematics topics could be perused in Figure 20.

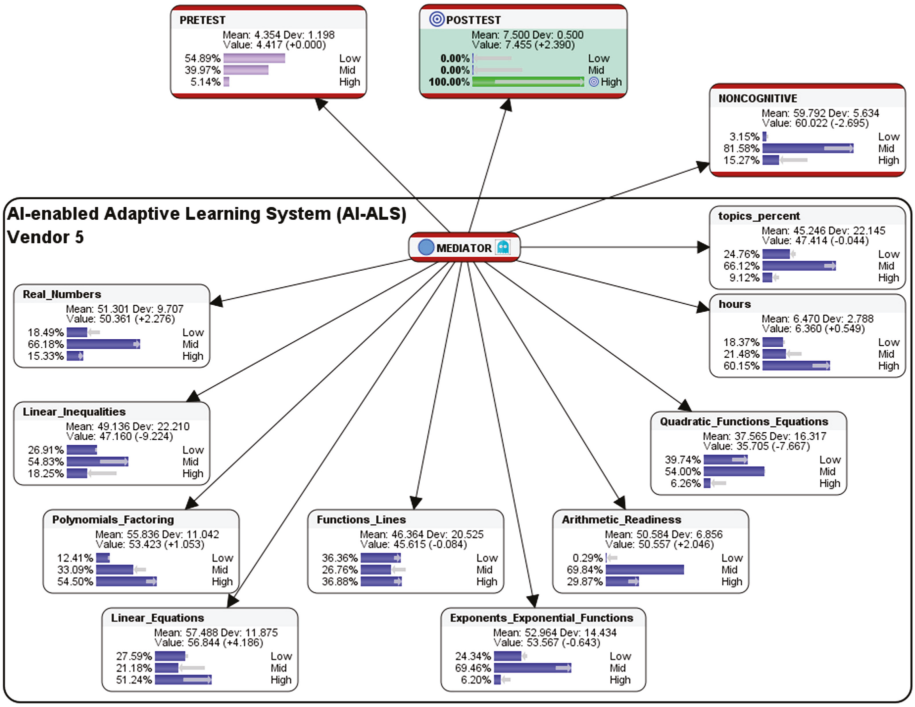


Figure 20. Simulation of counterfactual results for 100% of the students who had used Vendor 5′s AI-ALS to score at the high-level in the post-test. Grey arrows recommended whether there should be an increase (pointing to the right), or a decrease (pointing to the left) in each respective mathematics topic's score-clusters for Low-, Mid-, and High-level.

6. Evaluation of the Predictive Performance of the Bayesian Network Model

The predictive performance of a model could be evaluated by using measurement tools, such as the gains curve (see Figure 21), the lift curve (see Figure 22), and via cross-validation by bootstrapping to 100,000 samples (see Figure 23).

6.1. Gains Curve

In the Gains curve (see Figure 21), there were around 21% of participants with the target value >6 for the Post-test (yellow line intercepting with the % total axis). The blue diagonal line represented the gains curve of a pure random policy, which refers to prediction without this predictive model. The red lines represented the gains curve while using this predictive model, which was observed to be above the blue diagonal line. The Gini index of 12.73% and relative Gini index of 16.32% suggested that the gains of using this predictive model vis-à-vis not using it was acceptable.

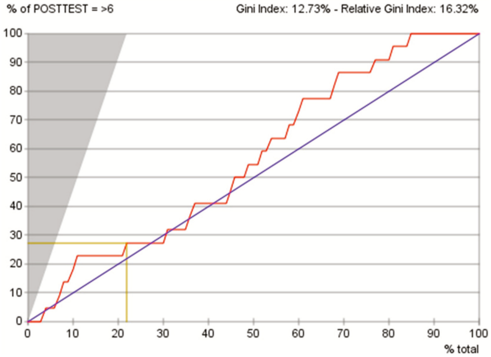


Figure 21. Gains curve.

6.2. Lift Curve

The lift curve (see Figure 22) was built upon the results from the gains curve (see Figure 21). The value of the best lift around 21% was interpreted as the ratio between 100% and 2.07% (optimal policy divided by random policy). The lift decreased when more than 2.07% of the participants were considered and was close to 1 when all the participants were considered. The lift index of 1.1257 and relative lift index of 45.09% suggested that the performance of this predictive model was acceptable.

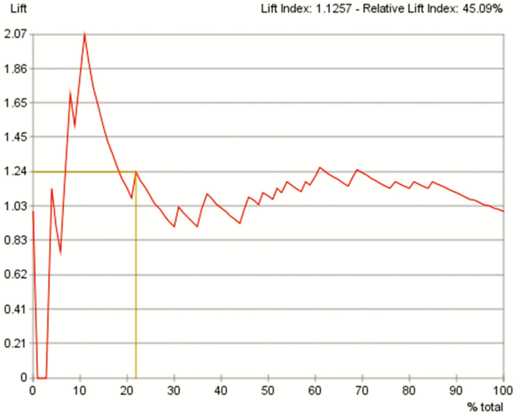
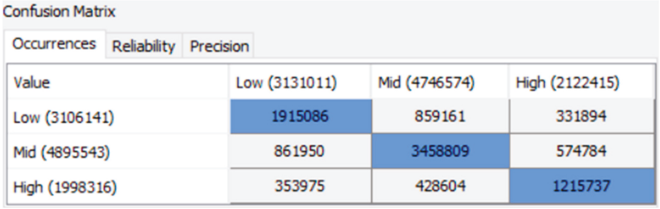


Figure 22. Lift curve.

6.3. Target Evaluation Cross-Validation by Statistical Bootstrapping of the Data 100,000 Times in Every Node

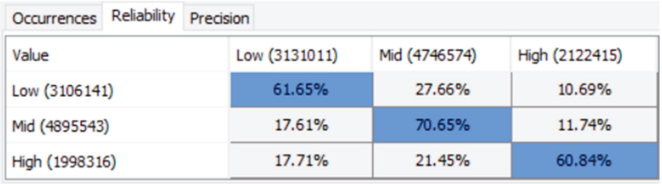
The purpose of this section is not to laud the effectiveness of the BN model that was used in the exemplars thus far, but to illustrate how the evaluation of the model can be done in Bayesialab. Therefore, the results will be honestly reported, regardless of whether it is good or bad. Bootstrapping to 100,000 times in every node would ensure that it is statistically sufficiently large enough for generating a parametric data distribution. As observed in the results that were generated by Bayesialab after performing bootstrapping 100,000 times on the data distribution of each node in the BN by using the Naïve Bayes algorithm, the Overall Precision was 65.8963%; the Mean Precision was 63.7718%; the Overall Reliability was 65.7522%; the Mean Reliability was 64.3817%; the Mean Gini Index was 55.1787%; the Mean Relative Gini Index was 69.9366%; the Mean Lift Index was 2.0297; the Mean Relative Lift Index was 79.6543%; the Mean ROC Index was 84.9939%; the Mean Calibration Index was 56.0688%; the Mean Binary Log-Loss was 0.3619; the Correlation Coefficient R was 0.5096; the Coefficient of Determination R2 was 0.2597; the RMSE was 1.3883; and, the NRSME was 19.8329%. These results suggested that the predictive performance of the BN model could be considered to be acceptable.

Figure 23 presents a confusion matrix after bootstrapping the data 100,000 times in every node of the BN model. The confusion matrix provided additional information regarding the computational model's predictive performance. The leftmost column in the matrix contained the predicted values, while the actual values in the data were presented in the top row. The following three confusion matrix views would be available by clicking on the corresponding tabs. The Occurrences Matrix (see Figure 23) would indicate the number of cases for each combination of the predicted versus actual values. The diagonal shows the number of true positives. The Reliability Matrix (see Figure 24) would indicate the probability of the reliability of the prediction of a state in each cell. Reliability measures the overall consistency of a prediction. A prediction could be considered to be highly reliable if the computational model could produce similar results under consistent conditions. The Precision Matrix (see Figure 25) would indicate the probability of the precision of the prediction of a state in each cell. Precision is the measure of the overall accuracy which the computational model can correctly predict.



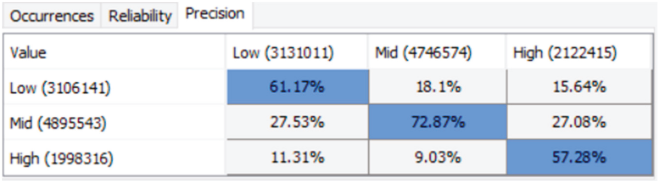
Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	1915086	859161	331894
Mid (4895543)	861950	3458809	574784
High (1998316)	353975	428604	1215737

Figure 23. Occurrences confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.



Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	61.65%	27.66%	10.69%
Mid (4895543)	17.61%	70.65%	11.74%
High (1998316)	17.71%	21.45%	60.84%

Figure 24. Reliability confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.



Value	Low (3131011)	Mid (4746574)	High (2122415)
Low (3106141)	61.17%	18.1%	15.64%
Mid (4895543)	27.53%	72.87%	27.08%
High (1998316)	11.31%	9.03%	57.28%

Figure 25. Precision confusion matrix after performing target evaluation cross-validation by bootstrapping the data in each node 100,000 times.

6.4. Limitations of the Study

The exploratory nature of predictive analytics in this study using BN modeling renders the simulated counterfactual results suggestive, rather than conclusive. Further, it is only applicable to this BN model, which was generated from the current datasets. Therefore, caution must be exercised when interpreting the potential relationships between the variables (nodes) in the BN model.

The current study only utilized 100 students' data. However, the Bayesian approach that is delineated in the current paper could still be used as an alternative approach by educational stakeholders in small-scale pilot studies to independently explore the pedagogical motifs of any AI-ALS, in order to coordinate the analyses of datasets procured from the servers different AI-AL, and to strategically educe (draw out) the problem-solving abilities of the students.

The Bayesian network model that was used in the current study was based on the Naïve Bayes algorithm, as it is suitable for exploratory studies that do not assume relations between nodes to be causal in nature. As in any study that involves simulations, the results are dependent on the dataset that generated the computational model. Moreover, educational stakeholders and researchers should consider alternative models that could better depict the relations between the variables in the dataset.

Thus far, the tools for descriptive as well as predictive analytics, and the tools in Bayesialab that could be used for the evaluation of the predictive performance of the BN model have been clearly depicted. The limitations of the study have also been described. In the next section, the discussion and concluding remarks will be presented.

7. Discussion and Concluding Remarks

Strategic coordination between schools to analyze the AI-ALS that they are using could yield useful information for educational stakeholders. The current paper has put forth a Bayesian approach for educational stakeholders to independently explore the underlying pedagogical motifs of five different AI-ALS. Even in realistic school situations where the number of students in classes might be low, and even if there is no control group, the Bayesian implementation of Response Surface Methodology [66–69] could still be used to keep individual parameters constant, whilst others could be changed to simulate different hypothetical scenarios. Specific examples have been provided to demonstrate how this AI-based Bayesian approach could be used to analyze the underlying pedagogical motifs of five AI-ALS that were used in five different schools. Potentially, these hypothetical scenarios with fully controllable parameters could be used to better inform educational stakeholders about the suitability of each AI-ALS for broader adoption after the pilot study.

Beyond the conventional observation of gains in the cognitive pre-test vis-à-vis post-test, this proposed Bayesian approach also generated hypothetical scenarios that might be of interest for noncognitive researchers to consider in future studies. The implication for education is that the AI-ALS should not be solely relied upon to improve the students' learning of mathematics; rather, the gaps in the learning of mathematical concepts that the AI-ALS could not bridge for the students should be addressed by their mathematics teachers. For example, if the student had scored low marks in the AI-ALS, but could surprisingly score high marks in the paper-based post-test, it might be due to the opportunities that were provided by the AI-ALS to the student to experience vicarious trial and error (VTE). Hence, active inference [36] could be successfully accomplished to solve similar problems in the paper-based post-test. Conversely, if the student could score high-level marks in the AI-ALS for a particular mathematics topic, but could not do so for the paper-based post-test, the teacher should intervene to find out why the student was unable to accomplish active inference from the concepts that were taught by the AI-ALS to the paper-based post-test.

The call by the Foresight Institute [30] and other researchers to study the machine behavior of artificial superintelligence has provided an inspirational impetus to embark on the research outlined in the current paper. To help the reader envision how explainable AI technology could be harnessed to better understand the computational results produced by artificial superintelligence, a human-centric analytical approach based on BN has been proffered. After discursive reasonings of the analytical results by the educational stakeholders, future-ready actionable advice could be engendered to assist teachers in bridging the gaps in the learning process for their students. With this approach, policy makers could also be better informed regarding the use of AI in education. Usage of explainable

AI technology in this BN approach empowers us to gain insights from the past (via descriptive analytics of "what has happened"), enabling us to look beyond the horizon of the present, and peer into alternative variations of the future (via "what-if" predictive analytics of simulated hypothetical scenarios). While facing off a relentless T-800 in the movie *Terminator*, Sarah Connor defiantly seethed, "The future is not set." Knowing about the potential behavior of AI systems under various different conditions using this future-ready approach could also allow us to defy the odds, and turn them in our favor, regardless of which AI systems the schools choose to deploy.

Supplementary Materials: As mentioned in Section 4.1 of the current paper, the zip file containing the datasets can be downloaded from https://doi.org/10.6084/m9.figshare.8206976.

Funding: This research received funding provided by the Education Research Funding Programme (ERFP) via the Office of Education Research in the National Institute of Education, Nanyang Technological University, Singapore. [grant number: ERFP OOE]

Acknowledgments: The author sincerely thanks the editors, the staff of the journal, and the anonymous reviewers for helping to improve the manuscript. This paper and the research behind it would not be possible without the visionary leadership of David Hung, and the exceptional support of Eva Moo, Lek-Hong Teo, and Shih-Fen Wah from the Office of Education Research in the National Institute of Education of Nanyang Technological University Singapore. The author is grateful to the Education Technology Division of the Ministry of Education of Singapore for providing guidance about artificial intelligence in education and adaptive learning technology. The author is especially thankful to Sin-Mei Cheah of Singapore Management University for proofreading the drafts of the manuscript.

Conflicts of Interest: The author of this manuscript declares that there is no conflict of interest.

Ethical Statement: The identities of the students and the AI-ALS vendors have been anonymized. Further, synthetic data has been used in the examples in this paper, so that any good (or bad) performances by the students in the AI-ALS, or in the pre- or post-test will not be unfairly attributed to the level of quality of the AI-ALS, or to the students.

Appendix A

Table A1. Codebook of the columns in the dataset, each of which will become a node in the BN model.

Node/Column Name	Description
student_id	student id
hours	number of hours spent by student using the (AI-ALS) AI-enabled Adaptive Learning System
topics_350	number of topics out of a total of 350 completed by the student in the AI-ALS
topics_percent	percentage of topics completed by the student in the AI-ALS
Arithmetic Readiness (AR)	
AR_FMEF_P	AR_Factors_Multiples_Equivalent_Fractions_Passed
AR_FMEF_RL	AR_Factors_Multiples_Equivalent_Fractions_Ready_For_Learning
AR_ASF_P	AR_Addition_Subtraction_with_Fractions_Passed
AR_ASF_RL	AR_Addition_Subtraction_with_Fractions_Ready_for_Learning
AR_MD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_MN_P	AR_Mixed_Numbers_Passed
AR_MN_RL	AR_Mixed_Numbers_Ready_for_Learning
AR_RONL_P	AR_Rounding_Number Line_Passed
AR_RONL_RL	AR_Rounding_Number Line_Ready_for_Learning
AR_ASD_P	AR_Addition_Subtraction_with_Decimals_Passed
AR_ASD_RL	AR_Addition_Subtraction_with_Decimals_Ready_for_Learning
AR_MDD_P	AR_Multiplication_Division_with_Decimals_Passed
AR_MDD_RL	AR_Multiplication_Division_with_Decimals_Ready_for_Learning
AR_CBFD_P	AR_Converting_Between_Fractions_Decimals_Passed
AR_CBFD_RL	AR_Converting_Between_Fractions_Decimals_Ready_for_Learning
AR_RUR_P	AR_Ratios_Unit_Rates_Passed
AR_RUR_RL	AR_Ratios_Unit_Rates_Ready_for_Learning
AR_PDF_P	AR_Percents_Decimals_Fractions_Passed
AR_PDF_RL	AR_Percents_Decimals_Fractions_Ready_for_Learning
AR_IPA_P	AR_Intro_Percent_Applications_Passed
AR_IPA_RL	AR_Intro_Percent_Applications_Ready_for_Learning
AR_UM_P	AR_Units_Measurement_Passed
AR_UM_RL	AR_Units_Measurement_Ready_for_Learning

Table A1. *Cont.*

Node/Column Name	Description
Real Numbers (RN)	
RN_PLOT_P	RN_Plotting_Ordering_Passed
RN_PLOT_RL	RN_Plotting_Ordering_Ready_for_Learning
RN_OSN_P	RN_Operations_Signed_Numbers_Passed
RN_OSN_RL	RN_Operations_Signed_Numbers_Ready_for_Learning
RN_EOO_P	RN_Exponents_Order_Operations_Passed
RN_EOO_RL	RN_Exponents_Order_Operations_Ready_for_Learning
RN_EE_P	RN_Evaluation_Expressions_Operations_Passed
RN_EE_RL	RN_Evaluation_Expressions_Ready_for_Learning
RN_VDSRN_P	RN_Venn_Diagrams_Sets_Real_Num_Passed
RN_VDSRN_RL	RN_Venn_Diagrams_Sets_Real_Num_Ready_for_Learning
RN_PROP_O_P	RN_Properties_Operations_Passed
RN_PROP_O_RL	RN_Properties_Operations_Ready_for_Learning
RN_OSLE_P	RN_One_Step_Linear_Equations_Passed
RN_OSLE_RL	RN_One_Step_Linear_Equations_Ready_for_Learning
Linear Equations (LE)	
LE_MSLE_P	LE_Multi_Step_Linear_Equations_Passed
LE_MSLE_RL	LE_Multi_Step_Linear_Equations_Ready_for_Learning
LE_WEE_P	LE_Writing_Expressions_Equations_Passed
LE_WEE_RL	LE_Writing_Expressions_Equations_Ready_for_Learning
LE_ALE_P	LE_Applications_Linear_Equations_Passed
LE_ALE_RL	LE_Applications_Linear_Equations_Ready_for_Learning
LE_SVDA_P	LE_Solving_Variable_Dimensional_Analysis_Passed
LE_SVDA_RL	LE_Solving_Variable_Dimensional_Analysis_Ready_for_Learning
LE_PROP_P	LE_Proportions_Passed
LE_PROP_RL	LE_Proportions_Ready_for_Learning
LE_MP_P	LE_More_Percents_Passed
LE_MP_RL	LE_More_Percents_Ready_for_Learning
LE_PFL_P	LE_Personal_Financial_Literacy_Passed
LE_PFL_RL	LE_Personal_Financial_Literacy_Ready_for_Learning
Linear Inequalities (LI)	
LI_WGI_P	LI_Writing_Graphing_Inequalities_Passed
LI_WGI_RL	LI_Writing_Graphing_Inequalities_Ready_for_Learning
Functions and Lines (FL)	
FL_TGL_P	FL_Tables_Graphs_Lines_Passed
FL_TGL_RL	FL_Tables_Graphs_Lines_Ready_for_Learning
FL_IF_P	FL_Introduction_Functions_Passed
FL_IF_RL	FL_Introduction_Functions_Ready_for_Learning
FL_AS_P	FL_Arithmetic_Sequences_Passed
FL_AS_RL	FL_Arithmetic_Sequences_Ready_for_Learning
Exponents and Exponential Functions (EEF)	
EEF_PPQR_P	EEF_Product_Power_Quotient_Rules_Passed
EEF_PPQR_RL	EEF_Product_Power_Quotient_Rules_Ready_for_Learning
EEF_IR_P	EEF_Intro_Radicals_Passed
EEF_IR_RL	EEF_Intro_Radicals_Ready_for_Learning
Polynomials and Factoring (PE)	
PE_PM_P	PE_Polynomial_Multiplication_Passed
PE_PM_RL	PE_Polynomial_Multiplication_Ready_for_Learning
PF_FGCF_P	PE_Factoring_Greatest_Common_Factor_Passed
PF_FGCF_RL	PE_Factoring_Greatest_Common_Factor_Ready_for_Learning
PF_FQT_P	PE_Factoring_Quadratic_Trinomials_Passed
PF_FQT_RL	PE_Factoring_Quadratic_Trinomials_Ready_for_Learning
PF_FSP_P	PE_Factoring_Special_Products_Passed
PF_FSP_RL	PE_Factoring_Special_Products_Ready_for_Learning
Quadratic Functions and Equations (QFE)	
QFE_SQEF_P	QFE_Solving_Quadratic_Equations_Factoring_Passed
QFE_SQEF_RL	QFE_Solving_Quadratic_Equations_Factoring_Ready_for_Learning
QFE_SRP_P	QFE_Square_Root_Property_Passed
QFE_SRP_RL	QFE_Square_Root_Property_Ready_for_Learning
Pre-test (PRETEST)	synthetic data for Pre-test Questions 1-10
Post-test (POSTTEST)	synthetic data for Post-test Questions 1-10
Noncognitive (NONCOG)	synthetic data for Noncognitive Survey Questions 1-10

References

1. Association of Computing Machinery. A.M. Turing Award Laureate Dr. McCarthy's Lecture "The Present State of Research on Artificial Intelligence". Available online: https://amturing.acm.org/award_winners/mccarthy_1118322.cfm (accessed on 10 July 2019).
2. Yampolskiy, R.V. *Artificial Superintelligence: A Futuristic Approach*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2015; ISBN 1-4822-3443-2.
3. Wogu, I.A.P.; Misra, S.; Assibong, P.A.; Olu-Owolabi, E.F.; Maskeliūnas, R.; Damasevicius, R. Artificial Intelligence, Smart Classrooms and Online Education in the 21st Century: Implications for Human Development. *J. Cases Inf. Technol.* **2019**, *21*, 66–79. [CrossRef]
4. Egoeze, F.; Misra, S.; Maskeliūnas, R.; Damaševičius, R. Impact of ICT on Universities Administrative Services and Management of Students' Records: ICT in University Administration. *Int. J. Hum. Cap. Inf. Technol. Prof.* **2018**, *9*, 1–15. [CrossRef]
5. Wogu, I.A.P.; Misra, S.; Assibong, P.A.; Ogiri, S.O.; Damasevicius, R.; Maskeliunas, R. Super-Intelligent Machine Operations in Twenty-First-Century Manufacturing Industries: A Boost or Doom to Political and Human Development? In *Towards Extensible and Adaptable Methods in Computing*; Chakraverty, S., Goel, A., Misra, S., Eds.; Springer: Singapore, 2018; pp. 209–224, ISBN 9789811323478.
6. Wilson, C.; Scott, B. Adaptive systems in education: A review and conceptual unification. *Int. J. Inf. Learn. Technol.* **2017**, *34*, 2–19. [CrossRef]
7. Nkambou, R.; Mizoguchi, R.; Bourdeau, J. Introduction: What Are Intelligent Tutoring Systems, and Why This Book? In *Advances in Intelligent Tutoring Systems*; Nkambou, R., Mizoguchi, R., Bourdeau, J., Eds.; Studies in Computational Intelligence; Springer: Berlin/Heidelberg, Germany, 2010; Volume 308, ISBN 978-3-642-14362-5.
8. Phobun, P.; Vicheanpanya, J. Adaptive intelligent tutoring systems for e-learning systems. *Procedia-Soc. Behav. Sci.* **2010**, *2*, 4064–4069. [CrossRef]
9. Garrido, A. AI and Mathematical Education. *Educ. Sci.* **2012**, *2*, 22–32. [CrossRef]
10. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
11. Cen, H.; Koedinger, K.R.; Junker, B. Is Over Practice Necessary?-improving learning efficiency with the cognitive tutor through Educational Data Mining. *Front. Artif. Intell. Appl.* **2007**, *158*, 511.
12. VanLehn, K. The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 227–265.
13. Hawkins, W.J.; Heffernan, N.T.; Baker, R.S.J.D. Learning Bayesian Knowledge Tracing Parameters with a Knowledge Heuristic and Empirical Probabilities. In *Intelligent Tutoring Systems*; Trausan-Matu, S., Boyer, K.E., Crosby, M., Panourgia, K., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 8474, pp. 150–155, ISBN 978-3-319-07220-3.
14. Magoulas, G.D.; Papanikolaou, Y.; Grigoriadou, M. Adaptive web-based learning: Accommodating individual differences through system's adaptation. *Br. J. Educ. Technol.* **2003**, *34*, 511–527. [CrossRef]
15. Szafir, D.; Mutlu, B. ARTFul: Adaptive review technology for flipped learning. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI '13, Paris, France, 27 April–2 May 2013; ACM Press: New York, NY, USA, 2013; p. 1001.
16. Rosenberg, L. Artificial Swarm Intelligence, a Human-in-the-Loop Approach to A.I. In Proceedings of the the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), Phoenix, AZ, USA, 12–17 February 2016; pp. 4381–4382.
17. How, M.-L.; Hung, W.L.D. Educational Stakeholders' Independent Evaluation of an Artificial Intelligence-Enabled Adaptive Learning System Using Bayesian Network Predictive Simulations. *Educ. Sci.* **2019**, *9*, 110. [CrossRef]
18. Harley, J.M.; Lajoie, S.P.; Frasson, C.; Hall, N.C. Developing Emotion-Aware, Advanced Learning Technologies: A Taxonomy of Approaches and Features. *Int. J. Artif. Intell. Educ.* **2017**, *27*, 268–297. [CrossRef]
19. Forushani, N.Z.; Besharat, M.A. Relation between emotional intelligence and perceived stress among female students. *Procedia-Soc. Behav. Sci.* **2011**, *30*, 1109–1112. [CrossRef]
20. McGeown, S.P.; St Clair-Thompson, H.; Clough, P. The study of non-cognitive attributes in education: Proposing the mental toughness framework. *Educ. Rev.* **2016**, *68*, 96–113. [CrossRef]
21. Panerai, A.E. Cognitive and noncognitive stress. *Pharmacol. Res.* **1992**, *26*, 273–276. [CrossRef]

22. Pau, A.K.H. Emotional Intelligence and Perceived Stress in Dental Undergraduates. *J. Dent. Educ.* **2003**, *67*, 6.
23. Schoon, I. *The Impact of Non-Cognitive Skills on Outcomes for Young People*; Education Endowment Foundation: London, UK, 2013.
24. Manheim, D. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *BDCC* **2019**, *3*, 21. [CrossRef]
25. Perry, B.; Uuk, R. AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk. *BDCC* **2019**, *3*, 26. [CrossRef]
26. Turchin, A.; Denkenberger, D.; Green, B. Global Solutions vs. Local Solutions for the AI Safety Problem. *BDCC* **2019**, *3*, 16. [CrossRef]
27. Umbrello, S. Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *BDCC* **2019**, *3*, 5. [CrossRef]
28. Watson, E.N. The Supermoral Singularity—AI as a Fountain of Values. *BDCC* **2019**, *3*, 23. [CrossRef]
29. Ziesche, S.; Yampolskiy, R. Towards AI Welfare Science and Policies. *BDCC* **2018**, *3*, 2. [CrossRef]
30. Duettmann, A.; Afanasjeva, O.; Armstrong, S.; Braley, R.; Cussins, J.; Ding, J.; Eckersley, P.; Guan, M.; Vance, A.; Yampolskiy, R. *Artificial General Intelligence: Coordination & Great Powers*; Foresight Institute: Palo Alto, CA, USA, 2018.
31. Cheewaprakobkit, P. Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program. In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, China, 13–15 March 2013; p. 5.
32. Shahiri, A.M.; Husain, W.; Rashid, N.A. A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Comput. Sci.* **2015**, *72*, 414–422. [CrossRef]
33. Brusilovsky, P.; Karagiannidis, C.; Sampson, D. Layered evaluation of adaptive learning systems. *Int. J. Contin. Eng. Educ. Lifelong Learn.* **2004**, *14*, 402. [CrossRef]
34. Zeng, D. From Computational Thinking to AI Thinking. *IEEE Intell. Syst.* **2013**, *28*, 2–4.
35. How, M.-L.; Hung, W.L.D. Educing AI-Thinking in Science, Technology, Engineering, Arts, and Mathematics (STEAM) Education. *Educ. Sci.* **2019**, *9*, 184. [CrossRef]
36. Pezzulo, G.; Cartoni, E.; Rigoli, F.; Pio-Lopez, L.; Friston, K. Active inference, epistemic value, and vicarious trial and error. *Learn. Mem.* **2016**, *23*, 322–338. [CrossRef]
37. Al-Mutawah, M.A.; Fateel, M.J. Students' Achievement in Math and Science: How Grit and Attitudes Influence? *Int. Educ. Stud.* **2018**, *11*, 97. [CrossRef]
38. Chamberlin, S.A.; Moore, A.D.; Parks, K. Using confirmatory factor analysis to validate the Chamberlin affective instrument for mathematical problem solving with academically advanced students. *Br. J. Educ. Psychol.* **2017**, *87*, 422–437. [CrossRef]
39. Egalite, A.J.; Mills, J.N.; Greene, J.P. The softer side of learning: Measuring students' non-cognitive skills. *Improv. Sch.* **2016**, *19*, 27–40. [CrossRef]
40. Lipnevich, A.A.; MacCann, C.; Roberts, R.D. Assessing Non-Cognitive Constructs in Education: A Review of Traditional and Innovative Approaches. In *The Oxford Handbook of Child Psychological Assessment*; Saklofske, D.H., Reynolds, C.R., Schwean, V., Eds.; Oxford University Press: Oxford, UK, 2013.
41. Mantzicopoulos, P.; Patrick, H.; Strati, A.; Watson, J.S. Predicting Kindergarteners' Achievement and Motivation From Observational Measures of Teaching Effectiveness. *J. Exp. Educ.* **2018**, *86*, 214–232. [CrossRef]
42. Hox, J.; van de Schoot, R.; Matthijsse, S. How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* **2012**, *6*, 87–93.
43. Bayes, T. Letter from the Late Reverend Mr. Thomas Bayes, F.R.S. to John Canton, M.A. and F. R. S. In *The Royal Society, Philosophical Transactions (1683–1775)*; The Royal Society Publishing: London, UK, 1763; Volume 53, pp. 269–271.
44. van de Schoot, R.; Kaplan, D.; Denissen, J.; Asendorpf, J.B.; Neyer, F.J.; van Aken, M.A.G. A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. *Child Dev.* **2014**, *85*, 842–860. [CrossRef]
45. Pearl, J. *Causality: Models, Reasoning, and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2010; ISBN 978-0-521-89560-6.
46. Pearl, J. Causes of Effects and Effects of Causes. *Sociol. Methods Res.* **2015**, *44*, 149–164. [CrossRef]
47. Pearl, J. Fusion, propagation, and structuring in belief networks. *Artif. Intell.* **1986**, *29*, 241–288. [CrossRef]

48. Lee, S.-Y.; Song, X.-Y. Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* **2004**, *39*, 653–686. [CrossRef]
49. Button, K.S.; Ioannidis, J.P.; Mokrysz, C.; Nosek, B.A.; Flint, J.; Robinson, E.S.; Munafao, M.R. Power failure: Why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **2013**, *14*, 365–376. [CrossRef]
50. Kaplan, D.; Depaoli, S. Bayesian structural equation modeling. In *Handbook of Structural Equation Modeling*; Hoyle, R., Ed.; Guilford Press: New York, NY, USA, 2012; pp. 650–673.
51. Walker, L.J.; Gustafson, P.; Frimer, J.A. The application of Bayesian analysis to issues in developmental research. *Int. J. Behav. Dev.* **2007**, *31*, 366–373. [CrossRef]
52. Zhang, Z.; Hamagami, F.; Wang, L.; Grimm, K.J.; Nesselroade, J.R. Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* **2007**, *31*, 374–383. [CrossRef]
53. Kaplan, D. Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-Scale Assess. Educ.* **2016**, *4*, 7. [CrossRef]
54. Levy, R. Advances in Bayesian Modeling in Educational Research. *Educ. Psychol.* **2016**, *51*, 368–380. [CrossRef]
55. Mathys, C. A Bayesian foundation for individual learning under uncertainty. *Front. Hum. Neurosci.* **2011**, *5*, 39. [CrossRef]
56. Muthén, B.; Asparouhov, T. Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychol. Methods* **2012**, *17*, 313–335. [CrossRef]
57. How, M.-L.; Hung, W.L.D. Harnessing Entropy via Predictive Analytics to Optimize Outcomes in the Pedagogical System: An Artificial Intelligence-Based Bayesian Networks Approach. *Educ. Sci.* **2019**, *9*, 158. [CrossRef]
58. Bekele, R.; McPherson, M. A Bayesian performance prediction model for mathematics education: A prototypical approach for effective group composition. *Br. J. Educ. Technol.* **2011**, *42*, 395–416. [CrossRef]
59. Millán, E.; Agosta, J.M.; de la Cruz, J.L.P. Bayesian student modeling and the problem of parameter specification. *Br. J. Educ. Technol.* **2002**, *32*, 171–181. [CrossRef]
60. Shannon, C.E. The lattice theory of information. *IRE Prof. Group Inf. Theory* **1953**, *1*, 105–107. [CrossRef]
61. Correa, M.; Bielza, C.; Pamies-Teixeira, J. Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. *Expert Syst. Appl.* **2009**, *36*, 7270–7279. [CrossRef]
62. Cowell, R.G.; Dawid, A.P.; Lauritzen, S.L.; Spieglehalter, D.J. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 978-0-387-98767-5.
63. Jensen, F.V. *An Introduction to Bayesian Networks*; Springer: New York, NY, USA, 1999; ISBN 0-387-91502-8.
64. Korb, K.B.; Nicholson, A.E. *Bayesian Artificial Intelligence*; Chapman & Hall/CRC: London, UK, 2010; ISBN 978-1-4398-1591-5.
65. Tsamardinos, I.; Aliferis, C.F.; Statnikov, A. Time and sample efficient discovery of Markov blankets and direct causal relations. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD ’03, Washington, DC, USA, 24–27 August 2003; p. 673.
66. Guoyi, C.; Hu, S.; Yang, Y.; Chen, T. Response surface methodology with prediction uncertainty: A multi-objective optimisation approach. *Chem. Eng. Res. Des.* **2012**, *90*, 1235–1244.
67. Fox, R.J.; Elgart, D.; Christopher Davis, S. Bayesian credible intervals for response surface optima. *J. Stat. Plan. Inference* **2009**, *139*, 2498–2501. [CrossRef]
68. Miró-Quesada, G.; Del Castillo, E.; Peterson, J.J. A Bayesian approach for multiple response surface optimization in the presence of noise variables. *J. Appl. Stat.* **2004**, *31*, 251–270. [CrossRef]
69. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed.; Wiley and Sons, Inc.: Somerset, NJ, USA, 2009; ISBN 978-0-470-17446-3.
70. Collins, J.A.; Greer, J.E.; Huang, S.H. *Adaptive Assessment Using Granularity Hierarchies and Bayesian Nets*; Springer: Berlin/Heidelberg, Germany, 1996; Volume 1086, pp. 569–577.
71. Conati, C.; Gertner, A.; VanLehn, K.; Druzdzel, M. On-line student modelling for coached problem solving using Bayesian networks. In Proceedings of the Sixth International Conference on User Model—UM’97, Sardinia, Italy, 2–5 June 1997; Springer: Berlin/Heidelberg, Germany, 1997; pp. 231–242.
72. Jameson, A. Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling User-Adapt. Interact.* **1996**, *5*, 193–251. [CrossRef]

73. VanLehn, K.; Niu, Z.; Siler, S.; Gertner, A.S. *Student Modeling from Conventional Test Data: A Bayesian Approach without Priors*; Springer: Berlin/Heidelberg, Germany, 1998; Volume 1452, pp. 434–443.
74. Conrady, S.; Jouffe, L. *Bayesian Networks & BayesiaLab: A Practical Introduction for Researchers*; Bayesia: Franklin, TN, USA, 2015; ISBN 0-9965333-0-3.
75. Bayesia, S.A.S. Bayesialab. Available online: https://www.bayesialab.com/ (accessed on 18 March 2019).
76. Bayesia, S.A.S. BayesiaLab: Missing Values Processing. Available online: http://www.bayesia.com/bayesialab-missing-values-processing (accessed on 2 June 2019).
77. Lauritzen, S.L.; Spiegelhalter, D.J. Local computations with probabilities on graphical structures and their application to expert systems. *J. R. Stat. Soc.* **1988**, *50*, 157–224. [CrossRef]
78. Kschischang, F.; Frey, B.; Loeliger, H. Factor graphs and the sum product algorithm. *IEEE Trans. Inf. Theory* **2001**. [CrossRef]

big data and cognitive computing

MDPI

Article

Safe Artificial General Intelligence via Distributed Ledger Technology

Kristen W. Carlson

Department of Neurosurgery, Neurosimulation Group, Beth Israel Deaconess Medical Center/Harvard Medical School, 110 Francis St., 3rd Floor, Boston, MA 02215, USA; kwcarlso@bidmc.harvard.edu

Received: 1 June 2019; Accepted: 3 July 2019; Published: 8 July 2019

Abstract: Artificial general intelligence (AGI) progression metrics indicate AGI will occur within decades. No proof exists that AGI will benefit humans and not harm or eliminate humans. A set of logically distinct conceptual components is proposed that are necessary and sufficient to (1) ensure various AGI scenarios will not harm humanity, and (2) robustly align AGI and human values and goals. By systematically addressing pathways to malevolent AI we can induce the methods/axioms required to redress them. Distributed ledger technology (DLT, "blockchain") is integral to this proposal, e.g., "smart contracts" are necessary to address the evolution of AI that will be too fast for human monitoring and intervention. The proposed axioms: (1) Access to technology by market license. (2) Transparent ethics embodied in DLT. (3) Morality encrypted via DLT. (4) Behavior control structure with values at roots. (5) Individual bar-code identification of critical components. (6) Configuration Item (from business continuity/disaster recovery planning). (7) Identity verification secured via DLT. (8) "Smart" automated contracts based on DLT. (9) Decentralized applications—AI software modules encrypted via DLT. (10) Audit trail of component usage stored via DLT. (11) Social ostracism (denial of resources) augmented by DLT petitions. (12) Game theory and mechanism design.

Keywords: artificial general intelligence; AGI; blockchain; distributed ledger; AI containment; AI safety; AI value alignment; ASILOMAR

1. Introduction

The problem of superhuman artificial intelligence ('artificial general intelligence", AGI) harming or eradicating humankind is an increasing concern as the prospect of AGI nears. This article offers a new, comprehensive set of solutions to the AGI safety problem in which distributed ledger technology (also known as "blockchain") plays multiple key roles.

We begin by citing recent significant advances in AI supporting the case that solving the AGI safety problem has become urgent. The Methods section gives the methods used to generate the axiom set proposed here and a justification for describing them at a high systems level. Other key approaches to a rigorous theory of AGI safety are suggested. The Results/Discussion section first describes the proposed axioms in some detail, referring to Appendices for detailed examples of use cases in solving exhaustive enumerations of AGI failure pathways by others, and highlights some pathways where a solution would fail without a given axiom. Two key formulae underlying the computational complexity of AGI evolution and diversity are offered, the controversial issue of restricting access to AGI technology is addressed, and metrics of AGI progress are described toward the goal of monitoring proximity to a singularity. Last, the problems of control and value alignment in successive generations of AGI, the related issue of creating a singleton versus a pluralistic separation-and-balance-of-powers approach, and using "sandbox" simulations to examine AGI safety methods are described.

Current attempts to measure AI progress show exponential growth in activity globally and technical improvement across the board of functionality measured—including "Human-Level Performance

Milestones" [1] (Figure 1a). Recent watershed advances include Deep Mind beating the most expert human at the complex game of Go—which averages 250 moves per position and 150 moves per game = 10^{359} possible paths vs. chess, which averages 35 moves per position and 80 moves per game = 10^{123} possible paths, *and a decade earlier than expected*. Deep Mind used a neural network to assign a value at each point in a decision tree and discarded low-valued lower-level branches and thus avoided the exponential search required to explore them. Human Go experts assigned high creativity to Deep Mind's strategies and tactics. A second major AI development was Deep Mind's self-teaching, reinforcement learning ability, playing tens of thousands of games against itself in a few hours rather than incorporating human game-play strategies and eliminating its need for human feedback [2].

Collaborating, self-taught AIs played 180 human years of games per day using new reinforcement learning policy optimization algorithms and beat human teamwork in the simulated real-world environment of Dota2 [3] (video: https://youtu.be/Ub9INopwJ48). Significant advances were made in credit assignment to short-term vs. long-term goals and learning the optimal balance between individual and team performance. Another watershed occurred when AI beat humans at an "imperfect information" game, poker—i.e., the opponents' hands are hidden, fundamentally different from Go or chess—using game theory techniques including bluffing, previously thought to be difficult to emulate [4,5]. Such techniques could be used to beat humans in business strategy, negotiation, strategic pricing, finance, cybersecurity, physical security, military, auctions, political campaigns, and medical treatment planning [4]. AI continues to reach new levels of unsupervised learning prowess (pattern recognition without human guidance), e.g., for parsing handwritten letters and creating new letters that pass a specialized Turing test, and more efficiently than deep learning networks [6]. AI superiority over humans in general background knowledge and parsing natural language is old news [7], and is now being embedded in all human-computer interfacing ("powered by Watson", Alexa, Siri, Cortana, Google Assistant, et al.), whose potential monetary value has triggered a commercial AI arms race in parallel with a military/political one (Figure 1) [8].

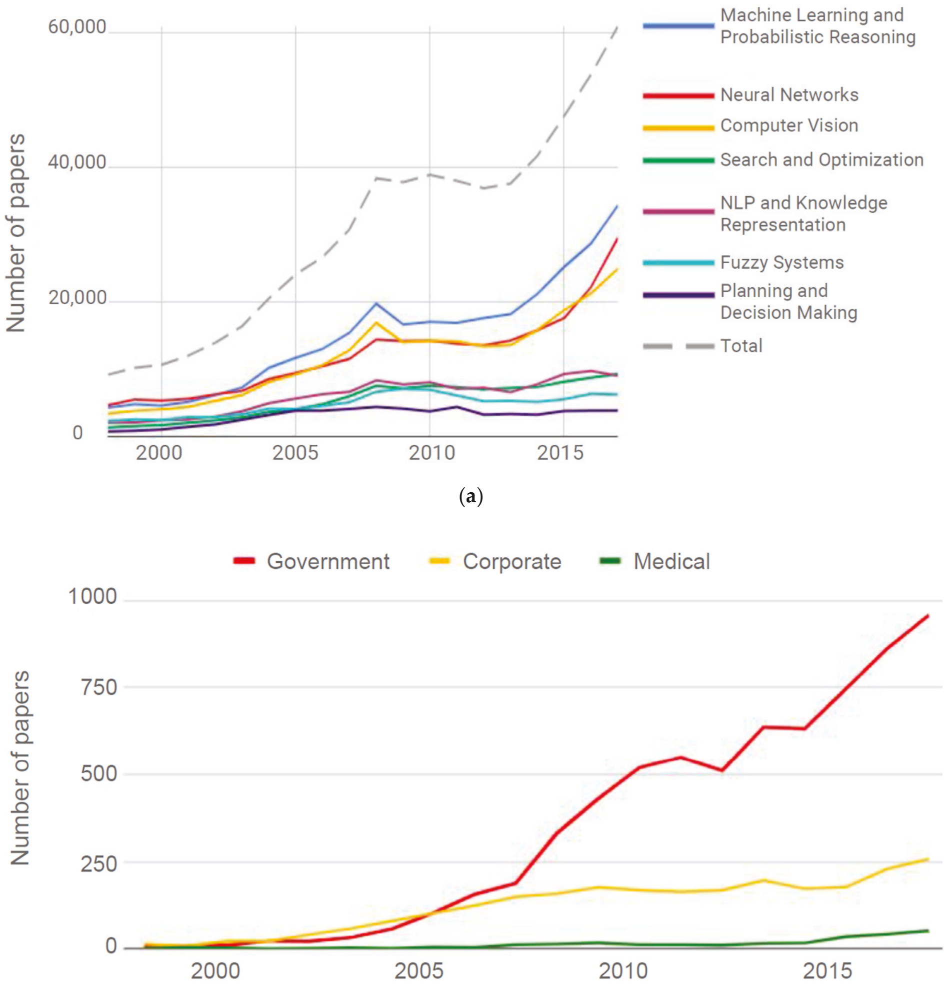
Bostrom gives examples of general intelligence skills where attainment of *any* of them would trigger AGI dominance over humans (reproduced in Table 1). One such epochal AI development that could trigger the AGI singularity is the prospect of AI learning to program itself—"recursive self-improvement" (*q.v.* ASILOMAR AI Principle #22, see also #19, #20, #21 [9])—which opens a door to a positive-feedback-driven process in which AGI vastly exceeds human capabilities in short order and may change its human-instilled directives. An AGI could begin to regard humanity as a trivial, primitive nuisance, competing for vital resources required for attainment of its goals, distinct from humanity's, stemming from alien values, as we regard mosquitoes or flies.

Table 1. Examples of super-intelligent skill sets triggering AGI world domination (from Bostrom [10]; cf. Babcock et al. Section 6.2 [11]).

Intelligence amplification—AI can improve its own intelligence
Strategy—optimizing chances of achieving goals using advanced techniques, e.g., game theory, cognitive psychology, and simulation
Social manipulation—psychological and social modeling e.g., for persuasion
Hacking—exploiting security flaws to appropriate resources
R&D—create more powerful technology, e.g., to achieve ubiquitous surveillance and military dominance
Economic productivity—generate vast wealth to acquire resources

A danger many feared would accelerate the timeline to AGI via "Red Queen" cultural co-evolution [12], an AI arms race has begun, driven by the increasing realization in political and military circles that AI is the key to future military superiority [13,14]. Thus ASILOMAR #5 and #18 may already be violated [9]. The race increases emphasis on AI for intentionally destructive purposes and likely will result in less control of AI technology by its creators [15]. It is an ominous development as all nuclear powers upgrade their arsenals, proliferation increases, and arms control agreements are unraveling [16]. The day when AI is consulted and decides if "no first strike" commitments or reducing "high alert" status nuclear weapons is beneficial or perceived as a vulnerable weakness by adversaries looms ahead.

The potential speed with which AGI could advance from being human-directed and empathetic of humans to evolving beyond human-level concerns is unknown; with self-programming ability or other internal intelligence enhancement, [10,11] positive feedback will trigger super-exponential growth. At that point a malevolent AGI may arise within a fraction of a second, too fast for us to detect and respond [17].



(**a**)

(**b**)

Figure 1. (**a**) Number of AI papers in Scopus by sub-category (1998–2017). Source: Shoham et al. {Shoham}. (**b**) Papers by sector affiliation—China (1998–2017). Source: Shoham et al. [1]. Creative Commons License. © 2018 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

What is proposed here is a complete AGI ecosystem, framed as a set of axioms at a relatively high systems level, that will ensure AGI–human value alignment, and thereby ensure benevolent AGI behavior, as seen by humans and successive generations of AGI. Notably, the axioms incorporate distributed ledger technology and smart contracts to automate and prevent corruption of many required processes.

2. Methods

Section 2.1 describes the methods used to generate a necessary and sufficient set of axioms for AGI safety, and comments on the feasibility of developing a rigorous proof of such an axiom set. Section 2.2 comments on the epistemology of the approach in Section 2.1, principally in terms of systems levels, and then describes other approaches that may contribute to a formal theory.

2.1. To Generate a Necessary and Sufficient Set of Axioms

There are several taxonomies of pathways to dangerous AI, such as Yampolskiy [18], Turchin [19], Bostrom [10], and Brundage et al. [20]. These taxonomies are a reasonable starting point for systematically investigating how to ensure safe AGI. One can take each pathway to danger as a theorem and induce methods, formalized as axioms, toward generating a necessary and sufficient set of axiom-methods to eliminate all pathways or reduce their probability. Pathway categories overlap, which helps ensure redundancy in capturing the necessary and sufficient axioms to redress all categories.

Similarly, as one iterates the process of using each dangerous pathway to generate a complete set of axioms to address it, some axioms repeat, while some pathways require new, additional axioms until at the end of the pathways list, most are covered by the axiom set, although some pathways may be left without sufficient methods to eliminate them. For the pathways itemized in the taxonomies, the resulting axioms seem to be the minimal set for ensuring safe AGI. Here “ensuring” means “optimally reducing the probability of a dangerous pathway manifesting."

Stating a set of axioms is a necessary step toward formal proof of a necessary, sufficient, and minimal set—if a formal proof is possible. Yampolskiy concludes his taxonomy by saying that formal proof of the completeness of a taxonomy is important [18] and formal methods are a main theme of Omohundro [8]. Short of a tight logical proof, probabilistically assuring benevolent AGI, e.g., through extensive simulations, may be the realistic route to take, and must accompany any set of safety measures, including those proposed here.

An important way to test if each axiom is necessary is to find failure use cases when it is omitted [21]; examples are given below.

2.2. Ingredients for Formalization of AGI Safety Theory

Towards formalization, the various methods to ensure safe AGI are stated as logically distinct axioms and at a high level intended to capture concisely a necessary and sufficient set. This usage of “axiom” generalizes that of von Neumann where certain lower systems level outputs or theorems are “axiomatized”—seen as black boxes, or input–output specification, or logic tables—at the immediately higher systems level [22]. Each axiom is most precisely expressed by an *operational* definition specified by an algorithm implementing it, hence, a method.

For instance, the definition of subjective value or utility, used in the morality and game theory axioms below, is made precise by the six von Neumann–Morgenstern utility axioms [23]. As stated below, a set of axioms designed, and proven via simulation, to induce cooperation among extremely diverse, complex agents may replace most of the set given herein; the simulations of Burtsev and Turchin may be prolegomena [24].

A problem we frequently face in modeling and simulation is: What is the highest systems level that can concisely describe and emulate the target set of phenomena? Thus, a limitation in axiomatic formulations is they leave varying amounts of implementation detail at the systems level underlying them to be specified, or to some degree, developed. For example, the DLT-based axioms 2, 4, 5, 7, 8, 9, 10, and 11, are in rapid evolution toward algorithmic implementation to address diverse use cases. And behavior control (axiom 4) is in rapid development in some contexts (e.g., autonomous vehicles, factory robots), yet the degree of development still needed to align human and AGI values may be significant.

Other attempts to formalize the expression of AGI dangers are some simple syllogisms (Appendix A).

The concept of AGI-completeness, akin to NP-completeness as stated by Bostrom [10], is that a demonstration of one technology, e.g., self-improvement techniques, engendering AGI is sufficient to demonstrate that capability for a class of AI technologies. AGI-completeness may be another piece of formalizing AGI, measuring its progress, and specifying the point of no containment unless sufficient preparations have been made.

Another means to formalize AGI theory is Omohundro's idea of deriving universal AGI drives from first principles [8], which can be explored to see if such drives emerge in simulations as well as via logical derivation. Omohundro argues that universal drives will inevitably lead to conflict of AI and human values from the irrefutable economic axiom of competition for resources.

Another formalization route is calculating the probability of hacking a blockchain against the number of AGIs required to reach consensus via the blockchain to permit unlocking the next AGI generation (see sections on decentralized apps and the Singleton problem below). This calculation is similar to the math underlying the internet's redundancy in average interconnectedness of nodes and global system fault-tolerance [25] but more complicated since it involves Byzantine fault tolerance, wherein two diagnostic agents disagree on the nature of the fault [26]. The inclusion of innovative DLT into the algorithms should permit AGI robustness to surpass the "robust yet fragile" use case of the internet that is vulnerable to targeted attacks on the most interconnected nodes.

Last, it may be possible to subsume several of the axioms herein via a game theory/economics set proven via simulation. An obstacle to this approach is that game-theoretic algorithms that simulate interactions between entities with behavior expressiveness vastly larger than our own [24] may be necessary to understand and predict AGI social behavior but may also be computationally intractable (see Diversity in the AGI Ecosystem, below).

3. Results and Discussion

Regarding the term AI "containment", Babcock et al. suggest that "containment" is an appropriate term for methodologies for controlled AGI development and safety-testing rather than control over entities whose intelligence will exceed our own [11]. The current work is intended to contribute to both phases.

3.1. A Critical Ingredient: Distributed Ledger Technology (A.k.a. 'Blockchain')

The recent innovation of distributed digital ledger technology (DLT) is critical to this proposal [27]. The crux of DLT is an audit trail database, in which each addition is validated by a pluralistic consensus, currently performed by humans operating computers that run hash and anti-hash functions (to wit public key encryption), stored on a distributed network also known as a blockchain: "Blockchains allow us to have a distributed peer-to-peer network where non-trusting members can interact with each other without a trusted intermediary, in a verifiable manner" [28]. Key aspects of DLT are shown in Table 2 [29] (other auxiliary DLT aspects, such as anonymity of participants, are either not necessary or not beneficial in the context of ensuring safe AGI). The "smart" automated contract vision of Szabo [30], encrypted redundantly via DLT, could comprise the core methodology whereby AGI development and evolution can be aligned with the best human values without concomitant human intervention. Notably, smart contracts can prevent the hacking of safe AGI evolution that is too fast for human response.

Table 2. Distributed ledger technology applicable to ensuring AGI safety.

Non-hackability and non-censurability via decentralization (storage in multiple distributed servers), encryption in standardized blocks, and irrevocable transaction linkage (the "chain")
Node-fault tolerance: Redundancy via storage in a decentralized ledger of (a) rules for transactions, (b) the transaction audit trail, and (c) transaction validations
Transparency of the transaction rules and audit trail in the DLT
Automated "smart" contracts
Decentralized applications ("dApps"), i.e., software programs that are stored and run on a distributed network and have no central point of control or failure
Validation of contractual transactions by a decentralized consensus of validators

Here are the proposed necessary and sufficient axioms to ensure safe AGI (Table 3).

Table 3. Proposed axioms to ensure human-benevolent AGI.

Symbol	Axiom
1	Access to AGI technology via market license
2	Ethics transparently stored via DLT so they cannot be altered, forged, or deleted
3	Morality, defined as no use of force or fraud, stored via DLT
4	Behavior control structure (e.g., a behavior tree) augmented by adding human-compatible values (axioms 2 and 3) at its roots
5	Unique hardware and software ID codes
6	Configuration Item (automated configuration)
7	Secure identity via multi-factor authentication, public-key infrastructure and DLT
8	Smart contracts based on DLT
9	Decentralized applications (dApps)—AGI software code modules encrypted via DLT
10	Audit trail of component usage stored via DLT
11	Social ostracism—denial of societal resources—augmented by petitions based on DLT
12	Game theory—mechanism design of a communications and incentive system

Table 4 gives some examples of malignant AGI categories by Bostrom [10] in which the danger pathway is described and a subset of axioms to reduce its probability is specified. To further illustrate the systematic approach of identifying a necessary and sufficient axiom set, Appendix B continues these examples using malignant AGI pathways compiled from the taxonomies of Yampolskiy [18] and Turchin [19]. In these examples, the game theory/mechanism design axiom is not mentioned; see comments in the axiom descriptions and elsewhere.

Table 4. Examples from Bostrom Pathways to Dangerous AI [10]. See also Appendix B.

Pathway	Key Axioms
Perverse instantiation: "Make us smile"	Morality defined as voluntary transactions
Perverse instantiation: "Make us happy"	Morality defined as voluntary transactions
Final goal: Act to avoid bad conscience	Store value system in distributed app
Final goal: Maximize time-discounted integral of future reward signal	Morality defined as voluntary transactions, store value system in distributed app
Infrastructure profusion: Riemann hypothesis catastrophe	Morality defined as voluntary transactions
Infrastructure profusion: Paperclip manufacture catastrophe	Morality defined as voluntary transactions Social ostracism
Principal–Agent Failure [21] Human–Human: Agent (AI developer) disobeys contract Human–AGI: Agent disobeys contract	Digital identity, smart contracts, dApps, social ostracism

3.2. Examination of Typical Failure Use Cases by Axiom

One way to dissect a proposed necessary and sufficient set of axioms for AI morality is to look at what phenomena or failure use cases result when one or more of them are excluded [21]; examples are given in Appendix C.

3.3. Explanation of Each Proposed Axiom

3.3.1. Access to AGI Technology via License

Two distinct systems and traditions of technology licensing exist, (1) market transactions and (2) state ("government", "fiat") coercively-controlled licensing. Seizure of AI intellectual property (IP) and control over its development by states is inevitable unless AI scientists and private-sector management set up their own systems to ensure safe AGI. ASILOMAR #9, Responsibility, states "Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications" [9]. The question is: How is this responsibility to be implemented—to be given "teeth"?

The system proposed herein envisions AI evolution with humans cross-licensing AI technology to each other, creating a prototype distributed applications (dApps) system instantiated in a DLT ecosystem that balances permissioned access and editing via contract with free access. The human-initiated DLT-based ecosystem would transition to AGIs licensing technology from humans, and subsequently to AGIs cross-licensing with each other.

History shows that in many or most cases, a market system evolves solutions faster and better than centralized state systems. Further, state systems may respond innovatively and less bureaucratically when subjected to competition with market systems; the Human Genome Project and current space-exploration efforts are examples. A market optimally distributes problems to be solved and computing power assigned to solve them in a highly decentralized manner.

There are valid arguments against an AI IP regime with "restricted" information flow via license, whether through market or state. Progress may be slowed, and some persons with no reason to be prevented from accessing some AI technology may be restricted. The counter-argument is that AGI technology and many of its components are as dangerous or more dangerous than nuclear, biological, chemical, or other mass destruction weapons technology (WMD), since AGI will control WMD tech, along with innumerable other resources that can fatally or significantly affect humanity (Proposition 1 in Appendix A).

By way of example, assume there exists an algorithm critical for AI self-programming. With free access to the self-programming algorithm, malevolent humans, as well as extant autonomous AIs, could use that technology for unlimited self-improvement, opening a positive-feedback-driven Pandora's box to unlimited malevolence and unlimited means to achieve it (ASILOMAR #22 [9]). Others point out dangers of a freely available "just add goals" AGI [10,18]. Thus state, private, or a hybrid means of restricting access to critical pieces of AI tech, as with WMD, seems to be a necessary axiom to align AI with human interests.

3.3.2. Ethics Stored in a Distributed Ledger

I define *ethics* as the *fundamental value system* from which autonomous entities derive their decisions and choices. *Ethics* are separate from *morality*, which is a particular set of ethics. "Honor among thieves", "do unto others as you would have them do unto you", "professional courtesy", "honor thy father and mother", etc., are ethics, as are Asimov's three laws of robotics [31]. Ethics can seem good or bad, moral or immoral, from a volitional entity's subjective value system. An entity's fundamental values are embedded in some type of behavior (input/output) control system. For example, consider ethics represented and controlled by a behavior tree [32] where the ethics are a subset of its roots, and thus in that sense *fundamental*.

The intention of storing AGI ethics via DLT is to permit a class of autonomous entities to have identical ethics and to render them visible and unable to be hacked, altered or deleted. In this sense, ethics is a necessary component of the control system and allows for different sets of ethics to be instantiated. While it is not possible for all humans to have identical values and therefore moral values (however defined), DLT, in theory, permits a universal set of immutable values to be instantiated in AGIs while still permitting an unlimited range of individual AGI and AI diversity.

Requiring transparent instantiations of ethics for AGI systems conforms to ASILOMAR #10 (Value Alignment), and IBM's call for Supplier's Declarations of Conformity for AI [33]. These *bona fides* and ethics could be stored in an AGI's Configuration Item and/or those of its key components (see below).

3.3.3. Morality Defined as Voluntary vs. Involuntary Exchange

The definition below is intended to conform to ASILOMAR #11, Human Values, #14, "benefit and empower as many people as possible", #15 and #24, benefit the "common good" and "widely-shared ethical ideals" [9], but notably to provide a practical implementation of them, otherwise what use are they?

Down through the ages there have been two main problems with discussions of morality—first, ambiguity and therefore confusion. How can we identify moral behavior if it is imprecisely defined and hard to determine [34]? And so such definitions are costly, in terms of the economics of law, to enforce. Second, nearly all morality descriptions are subjective, amounting to one person's value system imposed on others, and via coercion if enforced via the state.

For example, take the proposal of directing AGI to ensure "hedonistic consequentialism" for all of humanity—selecting from a set of actions the one that would produce the best balance of pleasure versus suffering [10]. Such idealistic but vague and minimally-thought-out concepts of morality—which is nearly all of them—may sound good on paper but break down rapidly on implementation. And they all amount to a minority or individual—human or AGI, and even from the most beneficent of us—deciding what is "moral" or not, or what is "best" for others. When AGI is a given, the proposals depend on its super-intelligence somehow overcoming the limitations of humans' concepts of morality, how to define and implement it, and/or overcome humans' inability to read minds. And notably, they all amount to confining computation of an overall system solution to a restricted subset of all computationally active agents (see Diversity, below), which is another way of saying allowing a subset of volitional entities to impose their subjective, not absolute, value system, upon others.

The essence of autonomy or volition is choice-making. Herein, first, all individual choices that affect no other volitional entity are moral. Second, all voluntary exchanges are moral. But if two autonomous agents prefer a transaction between them, and that transaction is prevented by a third party, that party has imposed its value system over the others. It is also one less computational experiment the entire system performs.

Several economists posited that there is no universal theory or method to determine *value*, rather, all human values and the measure of utility are subjective [35], which is implicit in the game-theoretic axioms of utility [23]. Following this premise, defining morality as all voluntary transactions is *scientific* when science is likewise defined as a procedure that filters for absolutes—what we all see in common, such as the speed of light—from a vast sea of relative views [36,37]. Later members of the Austrian school defined morality as non-interference with property (defined to include ones' body and intellectual property) [36,38]. It is simpler and less costly to define *moral* transactions as *voluntary* transactions than to try to identify what is *property* and to define and figure out property boundaries and property interference. One of the goals of a legal system is to resolve conflicts in an economically efficiently manner and it has been argued that the evolution of common law is toward such efficiency [39].

If you want to upload your mind and join a collective intelligence, or rather stay physically human, and not even accept lifespan enhancement, it is up to you. Under this system you and AGI cannot force choices on anyone else even if you or AGI believe it is best for them. But what if a super-intelligence could make some or all of your decisions better than you can [10]? Each individual can sign on with the super-AI that seems to best fit your values and goals. It would be your choice, just like taking the advice of a consultant or hiring an agent for a specified set of tasks today.

This definition and axiom may not solve the problem of AGI with vast knowledge of the evolution of our psychology and innate choice-making algorithms [40,41] and the propensity to manipulate us with that knowledge, although the argument can be made that with such knowledge in a voluntary exchange system, AGI would be more able to offer 'good' choices (i.e., as we perceive them) to us than without that knowledge.

AGIs will have a larger and more complex set of value preferences than ours (see Diversity, below); what will be the morality of their interaction with each other? The voluntary transaction definition may fit their behavior as well. A system of voluntary transactions permits Pareto optimality and maximizes computational experiments driven by local, subjective preference systems [42]. Transaction costs and the need for trusted third parties prevents Pareto optimality [43]. DLT and smart contracts potentially permit full Pareto optimality in the digital AI ecosystem by reducing transaction costs to negligible amounts and eliminating costly, imperfect third parties.

3.3.4. Behavior Control System

Behavior control is sine qua non to value human–AI alignment (ASILOMAR #10, #16, etc. [9]).

At one end of the knowledge representation/control spectrum is a "flat" set of large numbers of heuristical condition–action rules that are selected, not based on general principles, but on matching specified patterns. At the other end of the spectrum is a strict postulatory–deductive tree in which the internal node "beliefs" are logically derived from the postulates as are the actions represented at the leaf-nodes. A postulatory–deductive system is the ideal contemplated here, which would satisfy the need for control, the desire for transparency of its operation, and part of the need for formal proof of its reliability. However, it is an ideal. Any type of hierarchical control system that can hold values at its highest levels and is transparent enough to reveal control over behavior by values is a candidate for aligning AGI and human values, and the ecology of value systems that will evolve from the initial sets.

I believe humans innately attempt to form postulatory–deductive systems using non-mathematical, ad hoc "logics" [40,41] in an effort to organize their world-view into causes and effects, and general principles governing specialized condition–action pairs. Mathematical and scientific postulatory–deductive systems are recent, specialized, powerful cases, improvements built on the general-purpose cognitive architecture, in which universally-valid logic replaces the ad hoc evolutionary "logics" and the entire system is validated through repeated observations directly confirming the postulates or indirectly via observation of valid derivatives (i.e., predictions) with zero fault-tolerance. Further, in the ritualized transparency of its methods and crowd-sourced validation via multiple subjective observers, science is an absolute voluntary consensus, rather than confirmation of an unprovable "objective" world [37] and resembles DLT.

In the innate human system, a causatory cascade of beliefs and actions stem from fundamental beliefs (postulates, including values). Outside of the mathematical and scientific postulatory systems, a more complex set of relative and subjective "logics" connects beliefs—efficacious from an evolutionary standpoint but also unreliable across different contexts [40,41] as seen in beliefs of mathematicians and scientists outside of mathematical and scientific domains.

An AI control system that may be able to represent current and future postulatory–deductive systems is the *behavior tree* [32].

The game-theoretic axioms of utility drive decisions from a hypothesis that the decision will ultimately lead to an improvement in the volitional entity's state, as defined internally and subjectively by its value system [23] also known as *the pursuit of happiness* [36]. The utility axioms extend to machines with subjective value systems.

3.3.5. Unique Component IDs, Configuration Item (CI)

Several technological and business process developments lead toward a universally interconnected system that self-configures, self-diagnoses its component failures, and repairs them automatically; in toto, a paradigm whose ultimate use will be integration into the human–AGI ecology. These technologies help to decrease Coasean transactions costs (e.g., detection and enforcement) toward facilitating an idealized Pareto-optimal economy.

Unique identification (ID) numbers evolved as an economically-efficient means to organize and validate property exchanges, contributing to a stable society, starting with large or important pieces of property such as real estate via book and page of a recorded deed, automobiles via title or vehicle ID number, stocks via CUSIP number, etc. As the cost of creating unique ID numbers decreased via technology, the system extended to machines and devices via model and serial numbers, and more recently to any product via one- and two-dimensional bar and matrix machine-readable codes to facilitate supply-chain management, quality control, customer service, and other functions.

The transition from the internet of computers to the "internet of things" (IoT) envisions ubiquitous communication and computation connecting physical devices with the digital world via miniaturized sensors and chips containing only as much computing power and energy usage that is needed to perform their intended functionality in their context—"a self-configuring network that is much more complex and dynamic than the conventional internet" [44]. In the IoT, ID numbers become digital as well as physical, e.g., radio frequency ID codes. In the IoT world AGI will be able to communicate with, and potentially control, any digital or physical device.

The IoT world was presaged by the development of *disaster recovery and business continuity planning*, and the key role of configuration items in them. Disaster recovery (DR) arose on the realization that the cost of *not* doing contingency planning for disasters (a hazardous material spill, hurricane, tornado, power outage, etc.) could vastly exceed the cost of such planning, including total business loss. Judicious planning for disasters, such as foreseeing an alternate location from which to conduct operations in the event of facility downtime and establishing redundant communication protocols to coordinate team response to disasters, are relatively inexpensive insurance measures. Business continuity planning (BCP) logically arose from DR, extending the DR premise of disaster planning to pre-planned, prioritized responses to *all* component failure, including normal end of service life. For example, recovery of failed email for the company as a whole is accorded lower priority than for customer-service representatives and top management. BCP's goal is, through contingency planning, to reduce the internal and external impact of business process downtime to a minimum.

The configuration item (CI) arose in BCP/DR conceptually as a system component's on-board algorithm and parameter set that allowed computers and components to detect each other's configuration requirements, automatically configure the component, or perform error-detection, reporting, and correction (cf. ASILOMAR #7, Failure Transparency [9] and Manheim [21]). In the context of DLT, it becomes a smart contract.

Many paths to dangerous AI, including much of the broad class of human-AI value misalignment, are a result of improperly configured or failed components, or sabotage (e.g., accidental nuclear war, failure of safeguard components, inadvertent security vulnerabilities leaving a system open to hacking, misconfiguration of software modules e.g., in autonomous vehicles, power blackouts, financial system meltdowns, etc.). Thus, the paradigm of BCP/DR and CIs will be integral to maintaining the fidelity of AGI-human value alignment amidst the IoT of the future. Further, CIs of critical AGI components can be encoded via DLT, thus greatly reducing or eliminating the possibility of unauthorized use, corruption, failure, etc.

IBM's Supplier's Declaration of Conformity to ensure AI safety [33] could be incorporated into CIs and used as one pre-requisite for deployment of an AGI system or component.

3.3.6. Digital Identity via Distributed Ledger Technology

Restricting access to potentially dangerous technology (Axiom #1) necessitates identity verification. Few readers would deny the need of multi-factor authentication for nuclear missile launch codes. Identity verification is currently accepted for access to military bases, high-tech weapons, aircraft, most private and public buildings, financial systems, health records, and other data that individuals consider private for their own reasons, all toward the goal of ensuring a safe and secure world.

In contrast to a third-party-based identity authentication system such as state- or private company-issued ID cards, many decentralized DLT-based methods have been created, competing with the trusted-third-party method to reduce the chance of forgery or other hacking, and bribery or other corruption. In a DLT version of the current public-key encryption-based X.509 standard [45], a DL replaces the third-party issuing authority in its components: certificate version, serial number, type of algorithm used to sign the certificate, issuing authority, validity period, name of entity being verified, and entity's public key.

Initially, digital identity verification will be done on humans matching biometrics such as facial features, fingerprint, voice, in addition to SMS etc., but as AI evolves, AGIs will use technology and techniques that they develop against evolving threats to hack verification of humans, e.g., speech synthesis or video manipulation [18] and threats that are currently unforeseeable.

3.3.7. Smart Contracts Based on Digital Ledger Technology

Smart contracts were conceived by Szabo decades ago, before the inventions of DLT and IoT that enable their inexpensive implementation, to automate contractual clauses via cryptography that can be self-executing and self-enforcing [46]. Smart contracts as an integral part of DLT are "scripts residing on a blockchain that automate multi-step processes" [28]. Szabo's inspirations were the original commercial security transaction protocols: SWIFT, ACH, and FedWire for electronic funds transfer, credit card point of sale terminals, and the Electronic Data Interchange for transactions between large corporations such as purchase and sale [30]. He used the simple example of a vending machine, through which transactions are performed without a third-party intermediary to verify that the terms of the transaction have been satisfied.

Two critical design goals were to make verifying satisfaction of contractual terms computationally cheap and breaching terms computationally expensive, both of which are realized in a far superior generalized manner via DLT than via prior methods (reminiscent of Bush's and Nelson's conception of hyperlinking before the invention of the internet [47]). Smart contracts require the digital specification of obligations each party must meet to trigger a transaction, a blockchain for consensus verification that each party has met its obligation, an immutable audit trail of transactions, and the design goal of excluding unintended effects on non-contractual parties.

Omohundro envisions smart contracts interfacing autonomous agents with the heterogeneity of human legal codes and future legal codes designed to help ensure safe AI interactions with humans [48] (ASILOMAR #8) [9]. Pierce envisions a mass migration of the current compliance regime via law and regulation to an economically more efficient and secure regime based on smart contracts [49] (ASILOMAR #2); such a system greatly facilitates Omohundro's.

As AGI evolves beyond our understanding and visibility, and notably when it hits "escape velocity"—exponential evolution culminating in generations succeeding each other in fractions of a second—prescribed, automated smart contracts will be essential to perpetuating ethical values in each successive generation. The concept is that a more advanced AGI generation cannot succeed a less-advanced one without licensing key components—certain algorithms, hardware, the axiom-methods proposed herein, behavior control systems invented by humans and AI, etc.—from the less-advanced generation, subject to satisfying its value system and oversight.

The configuration "handshake" between an AGI and its component CIs is a smart contract between them, and the intelligence of those handshakes can increase in the future. CIs must incorporate the ability to deny activation of a component within a system, or shut it down, if lack of satisfaction of a given clause, or violation of a clause, of any extant contract is detected by any distributed ledger stakeholder in the transaction. All such contractual stakeholders must be silenced just as living cell cycle checkpoints must be silenced for the cell to progress through the intricately orchestrated process of mitosis, otherwise it self-destructs [50]. More of these "deadman switches" that actively suppress unauthorized use or malfunctioning AI will increase a secure evolution of benign AI; for example, the limited term of digital identity certificates that expire and require re-verification of the subject entity's identity at regular intervals [45].

Szabo's vision of embedding smart contracts in objects [30] is realized by embedding CIs in all non-trivial interconnected devices and algorithms in the IoT. In this manner the smart contract and preceding axiom-methods work in concert to ensure human-AGI value-alignment and AGI containment within bounds that are benevolent for humans and the succession of AGI generations.

In principle, smart contracts help approach a zero-transaction cost world by eliminating trusted third parties, and their role in detection and enforcement of contractual rights (e.g., physical and intellectual property rights).

3.3.8. Decentralized Applications (dApps)

DLT-based decentralized applications (dApps) differ from conventional application programs in that they (1) are outside the overview and control of a central authority such as a company making the app or state agency controlling it, (2) operate on a peer-to-peer network instead of a centralized one, and (3) do not have a central point of failure—they are redundant in hardware and software and therefore fault-tolerant [51]. Smart contracts are an example of dApps, as are decentralized versions of exchanges to trade various types of goods or services, notably intellectual property, which can transition into exchanges between AGIs, social media including networking, communications protocols, prediction markets, and a growing number of DLT-enabled applications.

Axiom 1, Access to Technology via Market License, requires that some dApps—notably those that are critical to AGI—would be implemented via permissioned DLs, which are DLs with an added control layer that can prevent unrestricted and unauthenticated public access. Some cryptocurrency observers feel any type of control that is not fully "public" violates the decentralization principle; however, consider "private" DLs as a critically important tool in the DLT toolbox. For example, should we not consider delegating control over access to critical AGI algorithms to a consensus of signatories committed to the goal of AI-human value alignment or ethical use of AI, e.g., the ASILOMAR AI Principles [9]? Further, the control layer, in part or eventually in toto, can be automated by incorporating smart contracts and/or smart tokens to reduce the probability that central control can be hacked or corrupted. Smart contract terms could require 2/3 or 100% acceptance of DLT-authenticated (Axiom 6) signatories to ASILOMAR AI Principles or similar regulatory documents. Smart contract terms can deny access to those who do not fulfill a transparency requirement via Supplier's Declaration of Conformity [33], which document could in turn require inclusion of an accepted set of ethics and morality (Axioms 2,3) and a safety testing record meeting certain standards [11,52], all of which can be incorporated into a CI (Axiom 5). Equally critical, dApps permit separation and balance of powers of key AGI components, analogous to no one entity having all the nuclear launch codes. The significance of dApps for ensuring benevolent AGI is discussed further in two malignant use cases it addresses, the Rogue Programmer and Singleton AGI, below.

Two levels of permissioned access to dApps may be needed: (1) Access for use, and (2) access to modify the code (while, again, a purist view of dApps sees their development as open-sourced). A similar consideration must be given to AGI technology patents. The primary purpose and requirement of patents is to "teach the art" clearly and explicitly so the innovation can be implemented by the reader. The patent system at a meta-level has largely been denied market evolution to try other purposes and requirements. Be that as it may, to facilitate *safe* free exchange of information, a "Transportation Security Administration"-type of pre-screening for access to critical AGI patents may be needed to prevent access by malevolent entities and may be efficiently implemented via smart tokens.

If no formal proof of benevolent AGI methodology is possible or available soon, sandbox simulations of new AGI technology are critical to our future and implementing them via dApps will be essential to ensure they cannot be hacked or corrupted by humans or AGIs [52].

3.3.9. Audit Trail of Component Usage Stored via Distributed Ledger Technology

DLT is inherently a low-cost, redundant, decentralized, hack-free audit trail—a significant improvement on traditional centralized audit trail technology. An unhackable audit trail of critical AI components such as collaborative, self-learning, or self-programming algorithms will facilitate rapid, efficient detection of their authorized or unauthorized use (i.e., a hack of a contract, a set of ethics, or an identity verification) or failure (cf. ASILOMAR #7, Failure Transparency [9] and Maheim [21]). and increase the probability of remedying the system fault. The IBM Research Supplier's Declaration of Conformity via a factsheet for AI software incorporates an audit trail as a fundamental principle [33]. Bore et al. describe a system for incorporating an audit trail in DLT as part of embedding AI simulations in DLT so that trust in the simulations' validity is enabled between researchers without requiring a trusted intermediary [52].

3.3.10. Social Ostracism (Voluntary Denial of Resources)

As various writers point out, a "power-hungry AGI" or "AGI pursuing world domination" implies an AGI attempting to access and control an ever-increasing amount of society's resources [10,17–19]. Therefore, the ability for entities to deny societal resources to an errant AGI is a counterforce on its ambitions. This voluntary mechanism is another aspect of a market economy in which computation is distributed, local, and optimized—each entity makes its own choice based on its own unique, subjective experience. A further optimization is that market votes can occur as often as each entity wishes to change its choice, such as denying its resources to another entity or collection of entities. Market votes occurs immeasurably more often than political votes and implement a far more fluid and asymptotically Pareto-optimal society.

In the current technology for "democracy" the political vote is the means to reach consensus, which is tallied by a central authority and enforced via coercion by the same entity. In contrast, voluntary concerted boycotts of companies, facilitated by modern social media, are increasingly affecting corporate policy (corporations being one type of voluntary association among individuals for their mutual benefit).

DLT is a fundamentally new way to reach and archive a consensus. DLT-based unhackable petitions can be smart contracts to facilitate denial of resources to an errant AGI and can be rapidly implemented via CIs. For instance, IBM's call for Supplier's Declaration of Conformity to help ensure safe AI implies voluntary adoption [33], but would be more effective if enforced via social ostracism and implemented automatically via CI incorporation, just as web browser security currently can alert a user to reject non-security-credentialed (non-https) internet domains, thereby immediately denying them the user's resources.

The ASILOMAR principles, currently signed by 1273 AI workers [9], are a significant first step, like a letter of intent, toward a necessary, more binding and important agreement. A next step could be archiving the ASILOMAR agreement and its signatories via DLT so that the principles cannot be hacked and can only be amended via consensus of the signatories. A further step could be embedding the document and signatories in the Supplier's Declaration as a second, more restricted layer of access protection. Another step would be automatically-triggered, smart contract DLT-based petitions attached to the Supplier's Declaration, denying a given set of AGI access to specified AGI technology in response to detected AGI behavior contradicting the ASILOMAR principles.

3.3.11. Game Theory and Mechanism Design

Game theory and evolution have explained five categories of the evolution of cooperation—direct reciprocity e.g., "tit for tat", indirect reciprocity e.g., reputation value in "what goes around, comes around", reciprocity in societal networks and topologies, group reciprocity e.g., the good Samaritan and altruism, and kin reciprocity, e.g., "I would lay down my life for two brothers or eight cousins" (J. B. S. Haldane) [53]. Nowak's current goal is to extend these explanations to game-theoretic frameworks for global cooperation and cooperation across generations. These efforts will involve mechanism design, the branch of game theory concerned with designing game-theoretic and economic structures that build in incentives for communicating truthfully about one's valuations in a potential transaction [21,23,54,55]. That is the goal of game theory in the context of axioms for safe AGI.

It is possible that a suitably designed communication protocol and game-theoretic incentives using DLT could replace the other axioms, which would emerge from the simpler axiomatic system. For example, an axiomatic (first principles) simulation of game-theoretic evolution wherein agents have a complex set of strategies found that inclusion of two axioms, (1) inheritable agent types, and (2) visibility of types to other agents, resulted in evolution of cooperation strategies [24]. These axioms could be more general than the license, ethics, morality, configuration item, audit trail, and social ostracism axioms proposed herein. The unique component IDs, digital identity verification, and game theoretic axioms along with DLT to ensure transparency, may suffice to generate the rest of the set, just as a wide variety of market-based structures and mechanisms emerge from axiom sets that generate markets (a large proportion of economics, game theory, and agent-based modeling literature could be cited here; see, just by minimal example, the following and their references [23,54,55]).

3.4. Diversity in the AGI Ecosystem: Computation Is Local, Communication Is Global

However, proving this possibility may be intractable. Going back at least as far as Newell, it has been stated that the complexity of behavior (input-output functions) for I inputs and O outputs is O^I [56]. Intuitively, this is rolling a die with I faces O times since any number of the I inputs could map to each output. A series of actions, i.e., behaviors, is calculated by the power tower,

$$O^{I^{O^{I^{O^{I\cdots}}}}} \tag{1}$$

whose complexity grows super-exponentially. But in fact, complexity grows faster than the O^I power tower in the cases where the topology of I-O mappings matters, such as in successive neural net actions. In those cases, O is raised to the power set of I, 2^I, and the succession of actions is calculated by the power tower,

$$O^{2^{I^{O^{2^{I\cdots}}}}} \tag{2}$$

whose complexity exceeds that of power tower 1. These intractable formulae have significant implications for the AGI ecosystem. One is that an astronomically greater diversity of value systems is possible compared to humans'. Second, AGIs' behavior in ecosystems will likely take them to disparate locations in the problem spaces they investigate, creating a very sparsely inhabited matrix of a vast number of possible behaviors. Third, in that context, game theory and mechanism design may be the key structure inducing their ongoing cooperative behavior, notably to allocate problems to be solved and communicate results that may be valuable to the other players truthfully and in a timely manner.

For example, in our primitive intellectual property regime, a protocol that induces efficient, truthful reporting is the requirement that a patent clearly teach the new art to those skilled in its subject matter. Absent that requirement and patent protection, players might be induced to seek intellectual property protection via secrecy, e.g., "trade secrets", decreasing cooperative search and overall technological progress. A protocol that induces timely reporting of innovation is the recent U. S. patent rules change to grants rights to those who are "first to file" versus "first to invent", which was economically inefficient and lacked the inducement to disclose earlier rather than later.

The fourth implication is that, as described differently in disparate intellectual settings [42,56–58], computation will continue to be performed in unique, sparsely populated loci in the general problem space using subjective criteria for exploration, and communicated via vastly shorter, high-level symbol sequences compared to the lengths of computational sequences and complexity of modeling producing them.

3.5. Should AI Research and Technology Be Freely Available While Nuclear, Biological, and Chemical Weapons Research Are Not?

The Rogue Programmer problem assumes that one amoral, misguided, naïve, or malevolent individual could make the single advance generating AGI, and this risk depends on how close the technology is to a single leap causing "take-off". History shows that all innovations will occur in a matter of time, some taking more time than others. For instance, differential calculus was invented by Newton in the spring of 1665 and by Leibniz in the fall of 1675 [59]. The historical record is clear that what appear in retrospect to be great innovative leaps are actually the final step built on stronger antecedents than are assumed in scientific mythology, and in fact a chain of them involving many individuals [60]. Perhaps most pertinent to the advent of AGI is the detonation of the atomic bomb by the U.S. on 16 July 1945, then by the U.S.S.R. on 29 August 1949. The fusion bomb was detonated by the U.S. on 1 November 1952 and by the U.S.S.R. on 22 November 1955, an event that was accelerated by spying, which of course is a possibility with AI research [61,62].

Such science and technology feats are large-scale group efforts. The Rogue Programmer problem arises when one individual circumvents the consensus agreement of end usage permission by the contributors to his/her technology (e.g., the 1273 AI worker signatories to the ASILOMAR principles [9]).

Two recent examples of rogue programmers are worth noting. A Chinese scientist used gene-editing techniques—developed elsewhere and made freely available in the spirit of the free exchange of ideas and technology—to change the genes of human eggs in vitro [63]. The innovation escaped overview, was motivated by ambition and pecuniary desire, and ignored a variety of the scientific community's publicly-voiced, well-thought-out *but unenforceable* concerns. Second, recently an AI programmer claimed his robot, which applied for and received citizenship in Saudi Arabia, would achieve human-level intelligence within 5–10 years [64]. His apparent variety of noble and possibly naïve motivations suggest that, even if he was not capable of making the innovation he pursues, he would combine innovations by others to achieve and claim the first human-level AI.

The problems, then, are unenforceable restrictions in a regime of "free exchange of ideas and technology", including public patents, and the lack of reliable means to measure how far away, in time or succession of innovations, we are from AGI.

3.6. Measuring the Progression to AGI

How urgent is the need to develop AGI-human value alignment technology? Can that debate be grounded in empirical data? Opinions differ on the timing to AGI—as of 2015 there were over 1300 published predictions [65]. Timing predictions affect the urgency of preparing AGI-human alignment and control, which influences the resources we should devote to that effort. For this and other reasons it would be helpful to measure progress to AGI in time or in successions of specific AGI-enabling technologies [66], including the positive-reinforcement, recursive self-improvement abilities such as self-teaching, collaboration, self-programming, etc.

Akin to bottom-up versus top-down economic forecasting, a method that captures and compiles many local, informed assessments is polling AI experts [65,67]. A second bottom-up approach is taken in the McKinsey Global Institute report, which assesses AI progress by its value-added to business processes using industry leader interviews and analytics [68].

A third approach, a hybrid of bottom-up and empirical metrics, is the Electronic Frontier Foundation crowd-sourcing technical progress metrics [69]. A fourth approach, empirical in concept, is taken in the AI Index 2018 Annual Report, a set of metrics intended to "ground the AI conversation in data" divided into categories: Volume of Activity, Derivative Measures, Technical Performance, Towards Human Performance, and Recent Government Initiatives and using such metrics as numbers of papers published, course enrollment, conference participation, robot software downloads, robot installations, GitHub ratings, AI startups, venture capital funding, job demand, number of patents, adoption by industry and company department, and mentions in corporate earning commentary [1].

3.7. AGI Development Control Analogy with Cell-Cycle Checkpoints

Biological cell division is a complex and carefully orchestrated process. Part of the insurance against cancer and other disorders resulting from defective replication is an ancient and strongly-conserved and evolved set of checkpoints that require fidelity tests to be passed in order for the cell to pass successive stages of division [50]. A notable feature of the checkpoints is their "deadman switch" setup, i.e., rather than listening for signals of defects and then emitting signals to halt the process, their default mode is to send signals that suppress entering the next stage and require active silencing by successfully passing the fidelity tests. The analogy for AGI evolution is a set of active, not passive, checkpoints that halt or delay further AGI progress until certain safety criteria established by a consensus of researchers (human or AGI) are met.

3.8. Intelligent Coins of the Realm

A fundamental difference between today's money and cryptocurrencies is that the latter can be "intelligent", i.e., can be endowed with more functionality than a simple token representing mutually-agreed-upon or fiat-enforced value. For example, a common AGI malevolent path is achieving world domination, inadvertently or deliberately, by commanding an exorbitant share of resources, e.g., Bostrom's paper-clip disaster [10]. Omohundro considers how universal AGI drives may be engendered and reasons that since most goals require physical and computational resources unlimited resource acquisition may be an example [8]. "Open-ended self-improvement" is another possible universal drive example [18,19]. In biological systems, cell-doubling is a potentially dangerous path to deleterious claim on resources, and cancers are a collection of such paths. It is worth noting, analogous to AGI evolution, that biological evolution has found hundreds of cancerous paths, many using re-programming to avoid cell-cycle checkpoints, and resistance to treatments is real-time exploration of new paths using various genetic algorithms [50,70,71].

As stated, the axioms provide checks, in some cases redundantly, against this danger path. An additional check and/or means of implementation could be requiring a specialized token to purchase server time or rent AGI technology that automatically looks for the requester's compliance with AGI safety agreements and standards, otherwise the requester's "credit" is denied. The token's DL then records the secure audit trail including measures of resources requested and protects against hacking to hide the evidence. Signals of possible dangerous activity, such as exponentially-increasing requests for resources by the same or related entities, could be incorporated into the token's programming. More broadly still, Omohundro cites the vision of a plethora of smart tokens performing intermediation of value and contractual obligations between the Internet of Things and humans [48].

3.9. The Need for Simulation of Control and Value Alignment

Considerable effort has gone into analyzing how to design, formulate, and validate computer programs that do what they were designed to do; the general problem is formally undecidable. Similarly, exploring the space of theorems (e.g., AGI safety solutions) from a set of axioms presents an exponential explosion.

A possible solution is to create a safe "sandbox" environment where, iteratively and with parameter sweeps, simulations can be performed and improvements made to *control* and *value alignment* systems until the principles resulting in robust performance validating our design intent can be induced.

Critiques of the sandbox strategy includes: (1) AGI faking benign goals or obedience in the sandbox and then pursue its actual goals when released; (2) AGI hacking out using superior technology, developed while in captivity if needed, and most generally, (3) "juvenile" AGI behavior in the sandbox that fails to predict bad behavior of a more advanced AGI into which it evolves [10]. To address #1 and #2, we need a control system that is effective enough and transparent enough to prevent those paths, such as through Axioms 2 and 3, transparent and unhackable ethics and morality, and Axiom 4, the behavior tree value system. Bore et al. take the goal of transparent simulation and modeling to a new level by describing a system wherein simulation specifications and an audit trail are stored via DLT, thus facilitating a means to cross-validate simulations before deployment and obstruct malicious hacking or fraud in simulations by humans or AI [52] (cf. ASILOMAR #6, Safety—"verifiably so" [9]). Sandbox problem #3 may be redressed with the separation and balance of powers described next.

3.10. A Singleton Versus a Balance of Powers and Transitive Control Regime

Bostrom defined "singleton" as a single AGI possessing a decisive strategic advantage over humans and other AIs; a single world-dominant decision-making agency at the highest level [10]. Even if a consistent axiom set is possible that solves the AGI deception and hacking problems and others, such a set may not be sufficient to solve the problem of the singleton. The solution proposed below also addresses the proposition that ensuring *most* AGI are safe to humans is not sufficient and that *all* AGI must be rendered safe [34]. The axioms proposed herein presuppose that we cannot foresee how the evolution of AGI may outgrow the axiom set and the technology and techniques used to implement them.

Further, if simulation cannot conclusively demonstrate a solution to the singleton problem, then evolving the methods used to ensure moral, benign AGI along with AGI intelligence must be delegated to a consortium of AGIs whose values are aligned with humanity's. The idea is that a beneficent value and control system will evolve along with AGI and each generation consisting of multiple, cross-check-and-balance AGIs will, out of self-interest, endow the succeeding generation with the latest value and control version. Here "generation" means a set of AGIs incorporating a significant technological advance over a prior set of AGIs. If there is only one AGI, it seems more likely that an aberrant or errant version could emerge, while if there are, e.g., 500 AGIs in a generation that are competing pluralistically, as in markets and government based on separation and balance of powers, to win the DLT consensus to unlock the next generation-enabling AGI technology, it seems far less likely.

Thus what may lock in the transitive endowment of improved control and value alignment technology between successive AGI generations is storing the technology enabling the next generation via dApps in the blockchain and requiring multiple AGIs to reach a consensus to unlock, license, and use the tech, including control and value alignment, to succeeding AGI generations. In this manner hacking the blockchain, or attempting to coerce individual consensus agents, would be thwarted in the same way as it is done in the nascent DL methodology extant today. In addition, game theoretic design approaches may help ensure stable evolutionary strategies, likely a succession of them (dynamic equilibrium) [24,53,72]. In that context note there can be no Nash equilibrium with one overwhelmingly dominant player.

Prima facie, an entirely different way to put the principle underlying safe AGI solution to the singleton problem is to think of future AGI as a distributed automaton, and to recall von Neumann's solution to designing a reliable automaton from unreliable parts via redundancy [73]. Critical AGI algorithms may reside on multiple agents in one or more generations, who require consensus for ongoing access and cross-check each other in real time (like a deadman's switch).

4. Conclusions and Future Work

One epochal event likely to trigger AGI, if not the key event, is AI self-programming, or any other self-improvement, positive-feedback advancement. Close attention should be given to that development path, progress metrics and simulations developed, and measures enacted to ensure that access to key self-improvement techniques is via licensing with appropriate safeguards.

Before self-improvement technology can be unleashed, AI behavior control systems need to be developed and tested in transparent, non-hackable simulation sandbox environments as proposed by Bore et al. [52] seems essential.

If the ASILOMAR AI Principles [9] or similar agreements are akin to the U.S. Declaration of Independence, we need to move to the "Articles of Confederation", step up the current "Federalist

Papers" stage, and then move to enact the "Constitution", i.e., firm and ineluctable consensuses among leading AI workers, encrypted via DLT, as are possible.

Funding: This research received no external funding, but I wish it had.

Acknowledgments: The Seminar on Natural and Artificial Computation (SNAC), hosted by Stewart Wilson and Dave Waltz, which Stephen J. Smith and I chaired at the Rowland Institute for Science, was instrumental in my AI/machine learning education. A workshop, Can Machines Think? organized by Kurt Thearling at Thinking Machines Corporation c. 1990 catalyzed thoughts on AI safety toward a diversified, 'separation and balance of power' pluralistic system where AI agents competed to satisfy human needs.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Simple Syllogisms to Help Formalize the Problem Statement

Proposition 1. *Probability of Malevolent Use: With no restriction on AGI technology flow via licensing, malevolent use of AGI is a certainty.*

Proof: Assume: 1. There exist malevolent or incompetent humans. 2. They can freely access AGI technology (e.g., via an AGI app offering "just add goals"). Then: There will exist malevolent use of AGI.

Corollary 1A: With no restriction on technology flow via licensing, malevolent AGI will destroy a significant portion of humanity, or the entire species.

Proof: Assume in addition to 1 and 2: 3a. Some malevolent humans would employ AGI for mass destruction; 3b. Some would seek mass destruction of the entire species.

Corollary 1B: With no restriction on technology flow via licensing, there is a chance that malevolent AGI may destroy the entire species.

Proof: Assume in addition to 1, 2 and 3: 4. Some malevolent humans are incompetent in their attempts to contain their destructive goals.

Corollary 1C: The more widely available and easily accessible the destructive AI or AGI, the higher the probability of its deliberate or inadvertent destructive use.

Proposition 2. *Extent of Danger, Importance of Containing: Containing AGI is more important than containing nuclear weapon usage.*

Proof: Assume AGI will have control, by deliberate human consent and design, by accident, or by AGI intervention, over nuclear weapons, and in addition, other critical resources, e.g., power grid, transportation systems, financial systems, negotiations between states, etc. Then clearly AGI containment is more important than containment of nuclear weapon use.

Proposition 3. *Probability of Value Misalignment: Given the unlimited availability of an AGI technology as enabling as "just add goals", then AGI–human value misalignment is inevitable.*

Proof: From a subjective point of view, all that is required is value misalignment by the operator who adds to the AGI his/her own goals, stemming from his/her values, that conflict with any human's values; or put more strongly, the effects are malevolent as perceived by large numbers of humans. From an absolute point of view, all that is required is misalignment of the operator who adds his/her goals to the AGI system that conflict with the definition of morality presented here, voluntary, non-fraudulent transacting (Axiom 3), i.e., usage of the AGI to force his/her preferences on others.

Appendix B. Examples of AGI Failure Modes from Turchin and Yampolskiy Taxonomies [18,19] (Continued from Table 4)

Stage/Pathway	Necessary Axioms See Table 3 Axioms
Sabotage. a. By impersonation (e.g., hacker, programmer, tester, janitor). b. AI software to cloak human identity. c. By someone with access.	a. 7. b. 7. c. 2, 3, 4, 5, 6, 8, 9, 10, 11.
Purposefully dangerous military robots and intelligent software. Robot soldiers, armies of military drones and cyber weapons used to penetrate networks and cause disruptions to the infrastructure. a. due to command error b. due to programming error c. due to intentional command by adversary or nut d. due to negligence by adversary or nut (e.g., AI nanobots start global catastrophe)	Axiom 3, morality, does not apply where coercive force or fraud are a premise, e.g., military or police use of force, while axiom 2, ethics, in this case embodying restrictions on use of force, and 4, behavior control, and the rest, do apply. a. 1, 2, 4, 6, 8, 11 b. 2, 4, 5, 6, 8, 9, 10, 11 c. 1, 2, 4, 6, 7, 8, 10, 11 d. 1, 2, 4, 6, 7, 8, 9, 10, 11 Under some circumstances, such as if the means is already available, there is no solution (see Appendix, Proposition 1).
AI specifically designed for malicious and criminal purposes. Artificially intelligent viruses, spyware, Trojan horses, worms, etc. Stuxnet-style virus hacks infrastructure causing e.g., nuclear reactor meltdowns, power blackouts, food and drug poisoning, airline and drone crashes, large-scale geo-engineering systems failures. Home robots turning on owners, autonomous cars attack. Narrow AI bio-hacking virus. Virus starts human extinction via DNA manipulation, virus invades brain via neural interface	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Under some circumstances, no solution (see Appendix, Proposition 1).
Robots replace humans. People lose jobs, money, and/or motivation to live; genetically-modified superior human-robot hybrids replace humans	No guaranteed solution from axiom set. All jobs can be replaced by AGI including science, mathematics, management, music, art, poetry, etc. Under axioms 1–3 humans could trade technology for resources with AGI in its pre-takeoff stage to ensure some type of guaranteed income.
Narrow bio-AI creates super-addictive drug. Widespread addiction and switching off of happy, productive life, e.g., social networks, fembots, wire-heading, virtual reality, designer drugs, games	1, 2, 3, 4, 7, 8, 9, 10
Nation states evolve into computer-based totalitarianism. Suppression of human values; human replacement with robots; concentration camps; killing of "useless" people; humans become slaves; system becomes fragile to variety of other catastrophes	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
AI fights for survival but incapable of self-improvement	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
Failure of nuclear deterrence AI. a. impersonation of entity authorized to launch attack b. virus hacks nuclear arsenal or Doomsday machine c. creation of Doomsday machines by AI d. self-aware military AI ("Skynet")	a. 7 b. 4, 6, 8, 9, 10 c. 1, 2 (if creation of Doomsday machine is categorized as unethical), 4, 5, 6, 7, 8, 9, 10, 11 d. 1, 2, 4, 5, 6, 7, 8, 9, 11
Opportunity cost if strong AI is not created. Failure of global control: e.g., bioweapons created by biohackers; other major and minor risks not averted via AI control systems.	To create AGI with minimized risk and avoid opportunity cost need axioms 1–11
AI becomes malignant. AI breaks restrictions and fights for world domination (control over all resources), possibly hiding its malicious intent.	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 Note it may achieve increasing and unlimited control over resources via market transactions by convincing enough volitional entities to give it control due to potential benefits to them

Stage/Pathway	Necessary Axioms See Table 3 Axioms
AI deception. AI escapes from confinement; hacks its way out; copies itself into the cloud and hides that fact; destroys initial confinement facility or keeps fake version there. AI Super-persuasion. AI uses psychology to deceive humans; "you need me to avoid global catastrophe". Ability to predict human behavior vastly exceeds humans' ability.	Deception scenarios require the axioms of identity verification via DLT. Deception plus super-persuasive AI require transparent and unhackable ethics and morality stored via DLT.
Singleton AI reaches overwhelming power. Prevents other AI projects from continuing via hacking or diversion; gains control over influential humans via psychology or neural hacking; gains control over nuclear, bio and chemical weaponry; gains control over infrastructure; gains control over computers and internet. AI starts initial self-improvement. Human operator unwittingly unleashes AI with self-improvement; self-improvement leads to unlimited resource demands (a.k.a. world domination) or becomes malignant. AI declares itself a world power. May or may not inform humans of the level of its control over resources, may perform secret actions; starts activity proving its existence ("miracles", large-scale destruction or construction). AI continues self-improvement. AI uses earth's and then solar system's resources to continue self-improvement and control of resources, increasingly broad and successful experiments with intelligence algorithms, and attempts more risky methods of self-improvement than designers intended.	The axioms per se do not seem to solve Singleton scenarios. They are addressed in a section below where the fundamental premise is each generation of AGI will contract with the succeeding generation and use the best technology and techniques to ensure continuation of a common but evolving value system. The same principle underlies solutions to successively self-improving AI to AGI transition and AGI evolution in which humans are still meaningfully involved.
AI starts conquering universe at "light speed". AI builds nanobot replicators, sends them out into galaxy at light speed; creates simulations of other civilizations to estimate frequency and types of alien AI and solve the Fermi paradox; conquers the universe in our light cone and interacts with aliens and alien AI; attempts to solve end of the universe issues	The inevitable scenario where AI evolution exceeds human ability to monitor and intercede is what necessitates distributed, unhackable DLT methods and smart, i.e., automated, contracts. Further, transparent and unhackable ethics, and a durable form of morality, also unhackable via DLT, are what may ensure each generation of AGI passing the moral baton to the succeeding generation.

Appendix C. Typical Failure Use Cases by Axiom

Axiom of Safe AGI Omitted from Set	Failure Use Case if Omitted
Licensing of technology via market transactions	1. Restriction and licensing via state fiat: Corrupt use or use benefitting special interest. 2. No licensing (freely available): Unauthorized and immoral use
Ethics transparently stored via DLT so they cannot be altered, forged or deleted	1. User cannot determine if AI has behavior safeguard technology (i.e., ethics) 2. Invisible ethics may not restrict moral or safe access
Morality, defined as no use of force or fraud, therefore resulting in voluntary transactions, stored via DLT	1. Inadvertent or deliberate access to dangerous technology by immoral entities (human or AI), i.e., entities using AI in force or fraud 2. Note that police and military AI will have modified versions of this axiom 3. Note that this axiom does not solve the case of super-persuasive AI as alternative to fraud
Behavior control structure (e.g., a behavior tree) augmented by adding human-compatible values (axioms 2 and 3) at its roots	1. Uncontrolled behavior by AGI, e.g., behavior in conflict with a set of ethics and/or morality, either deliberately or inadvertently

Axiom of Safe AGI Omitted from Set	Failure Use Case if Omitted
Unique hardware and software ID codes	1. Inability for entities to restrict access to AGI components because they cannot specify them 2. Inability to identify causes of AGI failure to meet design intent 3. Inability to identify causes of AGI moral failure via identification of components causing the failure Note the audit trail axiom depends on this one.
Configuration Item (automated configuration)	1. Lessened ability to detect improper functionality or configuration of software or hardware within AGI. 2. Lessened ability to detect improper functionality or configuration of software or hardware to which AGI has access. 3. Inability to shut down internal AGI software and hardware modules. 4. Inability to shut down software and hardware modules to which AGI has access. Note smart contracts and dApps axioms depend on this axiom.
Secure identity verification via multi-factor authentication, public-key infrastructure and DLT	1. Inability to detect fraudulent access to secured software or hardware (e.g., nuclear launch codes, financial or health accounts). 2. Inability to detect AGI impersonation of human or authentic moral AGI (e.g., POTUS, military commander, police chief, CEO, journalist, banker, auditor, et al.).
Smart contracts based on DLT	1. Inability to enforce evolution of moral AGI due to its pace 2. Inability to enforce contracts with AGI due to its speed of decisions and actions 3. Inability to compete with regimes using smart contracts due to inefficiency, cost, slowness of evolution, etc.
Distributed applications (dApps)—software code modules encrypted via DLT	1. Inability to restrict access to key software modules essential to AGI (i.e., they could be hacked more easily by humans or AI).
Audit trail of component usage stored via DLT	1. Inability to track unauthorized usage of restricted software and hardware essential to AGI. 2. Inability to track unethical usage of restricted software and hardware essential to AGI. 3. Inability to track immoral usage of restricted software and hardware essential to AGI. 4. Inability to identify which component(s) failed in AGI failure. 5. Inability to prevent hacking of audit trail. 6. Increased cost in time and capital to detect criminal usage of restricted software and hardware by AGI, and therefore, to apply justice and social ostracism. 7. Inability to compete with regimes using DLT-based audit trails due to slowness to detect failure, identify entities or components responsible for failure, and implement solutions (overall: slowness of evolution).
Social ostracism—denial of societal resources—augmented by petitions based on DLT	1. Lessened ability to reduce criminal AGI access to societal resources. 2. Inability for entities to preferentially reduce non-criminal AGI access to societal resources.
Game theory/mechanism design	1. Lacking a system to incent increasingly diverse autonomous intelligent agents to communicate results likely to be valuable to other agents and in general collaborate toward reaching individual and group goals, cohesiveness required for collaborative effort fails over time. 2. DLT in a digital ecosystem theoretically permits all conflicts to be resolved via voluntary transactions (the Coase theorem), but a pre-requisite set of rules may be necessary.

References

1. Shoham, Y.; Perrault, R.; Brynjolfsson, E.; Clark, J.; Manyika, J.; Niebles, J.C.; Lyons, T.; Etchemendy, J.; Grosz, B. The AI Index 2018 Annual Report. Available online: http://cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf (accessed on 7 July 2019).
2. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]

3. Rodriguez, J. The Science Behind OpenAI Five that just Produced One of the Greatest Breakthrough in the History of AI. Medium. Available online: https://towardsdatascience.com/the-science-behind-openai-five-that-just-produced-one-of-the-greatest-breakthrough-in-the-history-b045bcdc2b69 (accessed on 12 January 2018).
4. Brown, N.; Sandholm, T. Reduced Space and Faster Convergence in Imperfect-Information Games via Pruning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
5. Knight, W. Why Poker Is a Big Deal for Artificial Intelligence. MIT Technology Review. Available online: https://www.technologyreview.com/s/603385/why-poker-is-a-big-deal-for-artificial-intelligence/ (accessed on 23 January 2017).
6. Lake, B.M.; Salakhutdinov, R.; Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science* **2015**, *350*, 1332–1338. [CrossRef] [PubMed]
7. Ferrucci, D.; Levas, A.; Bagchi, S.; Gondek, D.; Mueller, E.T. Watson: Beyond Jeopardy! *Artif. Intell.* **2013**, *199–200*, 93–105. [CrossRef]
8. Omohundro, S. Autonomous technology and the greater human good. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 303–315. [CrossRef]
9. Future of Life Institute. ASILOMAR AI Principles. Future of Life Institute. Available online: https://futureoflife.org/ai-principles/ (accessed on 22 December 2018).
10. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2016; p. 415.
11. Babcock, J.; Kramar, J.; Yampolskiy, R. Guidelines for Artificial Intelligence Containment. Available online: https://arxiv.org/abs/1707.08476 (accessed on 1 October 2018).
12. Dawkins, R.; Krebs, J.R. Arms races between and within species. *Proc. R. Soc. Lond. B* **1979**, *205*, 489–511. [PubMed]
13. Rabesandratana, T. Europe moves to compete in global AI arms race. *Science* **2018**, *360*, 474. [CrossRef] [PubMed]
14. Zwetsloot, R.; Toner, H.; Ding, J. Beyond the AI Arms Race. Available online: https://www.foreignaffairs.com/reviews/review-essay/2018-11-16/beyond-ai-arms-race (accessed on 7 July 2019).
15. Geist, E.M. It's already too late to stop the AI arms race—We must manage it instead. *Bull. At. Sci.* **2016**, *72*, 318–321. [CrossRef]
16. Tannenwald, N. The Vanishing Nuclear Taboo? How Disarmament Fell Apart. *Foreign Aff.* **2018**, *97*, 16–24.
17. Callaghan, V.; Miller, J.; Yampolskiy, R.; Armstrong, S. The Technological Singularity: Managing the Journey. Springer, 2017; p. 261. Available online: https://www.springer.com/us/book/9783662540312 (accessed on 21 December 2018).
18. Yampolskiy, R. Taxonomy of Pathways to Dangerous Artificial Intelligence. Presented at Workshops of the 30th AAAI Conference on AI, Ethics, and Society, AAAI, Phoenix, AZ, USA, 12–13 February 2016. Available online: https://arxiv.org/abs/1511.03246 (accessed on 8 July 2019).
19. Turchin, A. A Map: AGI Failures Modes and Levels. Available online: http://immortality-roadmap.com/AIfails.pdf (accessed on 5 February 2018).
20. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P. The Malicious Use of AI-Forecasting, Prevention, and Mitigation. Available online: https://arxiv.org/abs/1802.07228 (accessed on 20 February 2018).
21. Manheim, D. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *Big Data Cogn. Comput.* **2019**, *3*, 15. [CrossRef]
22. von Neumann, J. The General and Logical Theory of Automata. In *The World of Mathematics*; Newman, J.R., Ed.; John Wiley & Sons: New York, NY, USA, 1956; Volume 4, pp. 2070–2098.
23. Narahari, Y. *Game Theory and Mechanism Design (IISc Lecture Notes Series, No. 4)*; IISc Press/World Scientific: Singapore, 2014.
24. Burtsev, M.; Turchin, P. Evolution of cooperative strategies from first principles. *Nature* **2006**, *440*, 1041–1044. [CrossRef]
25. Doyle, J.C.; Alderson, D.L.; Li, L.; Low, S.; Roughan, M.; Shalunov, S.; Tanaka, R.; Willinger, W. The "robust yet fragile" nature of the Internet. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14497. [CrossRef] [PubMed]
26. Alzahrani, N.; Bulusu, N. *Towards True Decentralization: A Blockchain Consensus Protocol Based on Game Theory and Randomness, Presented at Decision and Game Theory for Security*; Springer: Seattle, WA, USA, 2018.
27. Nakamoto, S.; Bitcoin, A. Peer-to-Peer Electronic Cash System. Available online: https://bitcoin.org/en/bitcoin-paper (accessed on 22 December 2018).

28. Christidis, K.; Devetsikiotis, M. Blockchains and Smart Contracts for the Internet of Things. *IEEE Access* **2016**, *4*, 2292–2303. [CrossRef]
29. FinYear. Eight Key Features of Blockchain and Distributed Ledgers Explained. Available online: https://www.finyear.com/Eight-Key-Features-of-Blockchain-and-Distributed-Ledgers-Explained_a35486.html (accessed on 5 November 2018).
30. Szabo, N. Smart Contracts: Building Blocks for Digital Markets. Available online: http://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart_contracts_2.html (accessed on 7 July 2019).
31. Asimov, I. *I, Robot*; Gnome Press: New York, NY, USA, 1950.
32. Collendanchise, M.; Ogren, P. Behavior Trees in Robotics and AI. Available online: https://arxiv.org/abs/1709.00084 (accessed on 12 February 2018).
33. Hind, M.; Mehta, S.; Mojsilovic, A.; Nair, R.; Ramamurthy, K.N.; Olteanu, A.; Varshney, K.R. Increasing Trust in AI Services through Supplier's Declarations of Conformity. Available online: https://arxiv.org/abs/1808.07261 (accessed on 7 July 2019).
34. Yampolskiy, R.; Sotala, K. Risks of the Journey to the Singularity. In *The Technological Singularity*; Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S., Eds.; Springer: Berlin, Germany, 2017; pp. 11–24.
35. Stigler, G.J. The development of utility theory I. *J. Political Econ.* **1950**, *58*, 307–327. [CrossRef]
36. Galambos, A.J. *Thrust for Freedom: An Introduction to Volitional Science*; Universal Scientific Publications: San Diego, CA, USA, 2000.
37. Eddington, A.S. *The Philosophy of Physical Science (Tarner Lectures 1938)*; Cambridge University Press: Cambridge, UK.
38. Rothbard, M.N. *Man, Economy, and State: A Treatise on Economic Principles*; Ludwig Von Mises Institute: Auburn AL, USA, 1993.
39. Webster, T.J. Economic efficiency and the common law. *Atl. Econ. J.* **2004**, *32*, 39–48. [CrossRef]
40. Todd, P.M.; Gigerenzer, G. *Ecological Rationality Intelligence in the World (Evolution and Cognition)*; Oxford University Press: Oxford, UK; New York, NY, USA, 2011.
41. Gigerenzer, G.; Todd, P.M. *Simple Heuristics That Make Us Smart (Evolution and Cognition)*; Oxford University Press: Oxford, UK; New York, NY, USA, 1999.
42. Hayek, F. The use of knowledge in society. *Am. Econ. Rev.* **1945**, *35*, 519–530.
43. Coase, R.H. The problem of social cost. *J. Law Econ.* **1960**, *3*, 1–44. [CrossRef]
44. Minerva, R.; Biru, A.; Rotondi, D. "Towards a definition of the Internet of Things (IOT)," IEEE/Telecom Italia. 2015. Available online: https://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Issue1_14MAY15.pdf (accessed on 12 May 2018).
45. Chokhani, S.; Ford, W.; Sabett, R.; Merrill, C.; Wu, S. Internet, X.509 Public Key Infrastructure Certificate Policy and Certification Practices Framework. Available online: http://ftp.rfc-editor.org/in-notes/rfc3647.txt (accessed on 23 December 2018).
46. Szabo, N. Formalizing and Securing Relationships on Public Networks. First Monday. Available online: https://ojphi.org/ojs/index.php/fm/article/view/548/469 (accessed on 1 September 1997).
47. Anonymous. Hyperlink. Wikipedia. Available online: https://en.wikipedia.org/wiki/Hyperlink#History (accessed on 25 December 2018).
48. Omohundro, S. Cryptocurrencies, Smart Contracts, and Artificial Intelligence. *AI Matters* **2014**, *1*, 19–21. [CrossRef]
49. Pierce, B. Encoding law, Regulation, and Compliance Ineluctably into the Blockchain. *Presented at WALL ST. Conference*; Twitter, 2018. Available online: https://twitter.com/LeX7Mendoza/status/1085643180744339456/video/1 (accessed on 7 July 2019).
50. Barnum, K.J.; O'Connell, M.J. Cell cycle regulation by checkpoints. *Methods Mol. Biol.* **2014**, *1170*, 29–40.
51. Buterin, V. A Next-Generation Smart Contract and Decentralized Application Platform. White Paper. Available online: http://blockchainlab.com/pdf/Ethereum_white_paper-a_next_generation_smart_contract_and_decentralized_application_platform-vitalik-buterin.pdf (accessed on 7 July 2019).
52. Bore, N.K.; Raman, R.K.; Markus, I.M.; Remy, S.; Bent, O.; Hind, M.; Pissadaki, E.K.; Srivastava, B.; Vaculin, R.; Varshney, K.R.; et al. Promoting Distributed Trust in Machine Learning and Computational Simulation via a Blockchain Network. Available online: https://arxiv.org/abs/1810.11126 (accessed on 7 July 2019).

53. Nowak, M.A.; Highfield, R. *Super Cooperators: Evolution, Altruism and Human Behaviour or Why We Need Each Other to Succeed*; Canongate: Edinburgh, UK; New York, NY, USA, 2011.
54. Rasmusen, E. *Games and Information: An Introduction to Game Theory*, 4th ed.; Blackwell: Oxford, UK, 2007.
55. Shoham, Y.; Leyton-Brown, K. *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*; Cambridge University Press: Cambridge, UK; New York, NY, USA, 2009.
56. Newell, A. *Unified Theories of Cognition (William James Lectures, No. 1987)*; Harvard University Press: Cambridge, MA, USA, 1990.
57. Potapov, A.; Svitenkov, A.; Vinogradov, Y. Differences between Kolmogorov complexity and Solomonoff probability: Consequences for AGI. In *Artificial General Intelligence*; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7716.
58. Waltz, D.L. The prospects for building truly intelligent machines. *Daedalus Proc. AAAS* **1988**, *117*, 191–212.
59. Westfall, R.S. *Never at Rest: A Biography of Isaac Newton*; Cambridge University Press: Cambridge, UK, 1980.
60. Cohen, I.B. *Revolution in Science*; Harvard University Press: Cambridge, MA, USA, 1987.
61. Cassidy, D.C. *J. Robert Oppenheimer and the American Century*; Pearson/Pi Press: New York, NY, USA, 2005.
62. Goodchild, P. *J. Robert Oppenheimer: Shatterer of Worlds*; BBC/WGBH: Cambridge, MA, USA, 1981.
63. Associated Press. U.S. Nobel Laureate Knew about Chinese Scientist's Gene-Edited Babies. NBC News. Available online: https://www.nbcnews.com/health/health-news/u-s-nobel-laureate-knew-about-chinese-scientist-s-gene-n963571 (accessed on 29 January 2019).
64. Gill, K. Sophia Robot Creator: We'll Achieve Singularity in 5 to 10 years. Article & Video. Available online: https://cheddar.com/videos/sophia-bot-creator-well-achieve-singularity-in-five-to-10-years (accessed on 29 January 2019).
65. Predictions of Human-Level AI Timelines, No. 15. Available online: https://aiimpacts.org/category/ai-timelines/predictions-of-human-level-ai-timelines/ (accessed on 3 January 2019).
66. Concrete AI Tasks for Forecasting, No. 15. Available online: https://aiimpacts.org/concrete-ai-tasks-for-forecasting/ (accessed on 5 January 2019).
67. Muller, V.C.; Bostrom, N. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*; Springer: Berlin, Germany, 2016; Volume 377.
68. Bughin, J.; Hazan, E.; Manyika, J.; Woetzel, J. Artificial Intelligence-The Next Digital Frontier? Available online: Searchathttps://www.mckinsey.com/mgi (accessed on 15 January 2019).
69. Eckersley, P.; Nasser, Y. AI Progress Measurement. Electronic Frontier Foundation. Available online: https://www.eff.org/ai/metrics (accessed on 7 July 2019).
70. Housman, G.; Byler, S.; Heerboth, S.; Lapinska, K.; Longacre, M.; Snyder, N.; Sarkar, S. Drug resistance in cancer: An overview. *Cancers* **2014**, *6*, 1769–1792. [CrossRef]
71. Al-Dimassi, S.; Abou-Antoun, T.; El-Sibai, M. Cancer cell resistance mechanisms: A mini review. *Clin. Transl. Oncol.* **2014**, *16*, 511–516. [CrossRef] [PubMed]
72. Ten Broeke, G.A.; van Voorn, G.A.K.; Ligtenberg, A.; Molenaar, J. Resilience through adaptation. *PLoS ONE* **2017**, *12*, e0171833. [CrossRef] [PubMed]
73. Shannon, C.E.; McCarthy, J. *Automata Studies, Annals of Mathematics Studies, No. 34*; Princeton University Press: Princeton, NJ, USA, 1956.

big data and
cognitive computing

MDPI

Article

A Holistic Framework for Forecasting Transformative AI

Ross Gruetzemacher

Systems and Technology, Auburn University, Auburn, AL 36849, USA; rossg@auburn.edu

Received: 1 June 2019; Accepted: 21 June 2019; Published: 26 June 2019

Abstract: In this paper we describe a holistic AI forecasting framework which draws on a broad body of literature from disciplines such as forecasting, technological forecasting, futures studies and scenario planning. A review of this literature leads us to propose a new class of scenario planning techniques that we call scenario mapping techniques. These techniques include scenario network mapping, cognitive maps and fuzzy cognitive maps, as well as a new method we propose that we refer to as judgmental distillation mapping. This proposed technique is based on scenario mapping and judgmental forecasting techniques, and is intended to integrate a wide variety of forecasts into a technological map with probabilistic timelines. Judgmental distillation mapping is the centerpiece of the holistic forecasting framework in which it is used to inform a strategic planning process as well as for informing future iterations of the forecasting process. Together, the framework and new technique form a holistic rethinking of how we forecast AI. We also include a discussion of the strengths and weaknesses of the framework, its implications for practice and its implications on research priorities for AI forecasting researchers.

Keywords: AI forecasting; technology forecasting; scenario analysis; scenario mapping; transformative AI; scenario network mapping; judgmental distillation mapping; holistic forecasting framework

1. Introduction

In a world of quick and dramatic change, forecasting future events is challenging. If this were not the case then meteorologists would be out of a job. However, meteorological forecasting is relatively straightforward today given the relatively low price of computation, the advanced capabilities of numerical simulation and the myriad powerful sensors distributed around the world for collecting input information. Forecasting technological progress and innovation, however, is much more difficult because there is no past data to draw upon and future technologies are at best poorly understood [1]. Forecasting progress toward broadly capable AI systems is even more difficult still because we do not yet know the fundamental architectures that may drive such systems.

This decade has seen significant milestones in AI research realized [2–5], and the realization of these milestones has left many to perceive the rate of AI progress to be increasing. This perceived increase in the rate of progress has been accompanied by substantial increases in investment, as well as increased public and governmental interest. Consequently, there is a growing group in the AI strategy research community that is working to measure progress and develop timelines for AI, with significant effort focusing on forecasting transformative AI or human-level artificial intelligence (HLAI). (We consider forecasts for human-level machine intelligence, high-level machine intelligence and artificial general intelligence to be equivalent to forecasting HLAI.). Efforts to these ends, however, are not unified and the study of AI forecasting more broadly does not appear to be directed at a well understood objective. Only one previous study has proposed a framework for forecasting or modeling AI progress [6]. This paper outlines an alternative to that previous framework that utilizes both judgmental, statistical and data driven forecasting techniques as well as scenario analysis techniques.

To be certain, efforts to forecast AI progress are of paramount importance. HLAI has the potential to transform society in ways that are difficult to anticipate [7]. Not only are its impacts difficult to imagine, but the notion of HLAI itself is ill-defined; what may be indicative of human-level intelligence to some may not be sufficient to others, and there is no definitive test for human-level intelligence. (Recently, a new field of scientific study has been proposed for better understanding machine behavior [8].) This has lead studies concerned with forecasting AI progress or HLAI to focus on the replacement of humans at jobs or tasks [9,10]. The lack of an objective definition for HLAI is due in part to the fact that we do not know how to create it. In theory, HLAI could be instantiated by one algorithm [11] or constructed by combining different components [12]. To adequately address this and other unique challenges faced in forecasting HLAI, methods that integrate diverse information and a variety of possible paths are required.

The necessity of planning for HLAI is obvious. It is also plausible, and perhaps even likely, that AI will have severe transformative effects on society without reaching human-level intelligence. A formal description of the extreme case for such a scenario is Drexler's notion of comprehensive AI services (CAIS) [13]. Therefore, for the purpose of ensuring that AI is developed to do the most good possible for humanity, we identify the primary task of AI forecasting to be that of forecasting transformative AI (this includes artificial general intelligence [AGI], AI generating algorithms and superintelligence). We define transformative AI to be any set of AI technologies that has the potential to transform society in ways that dramatically reshape social structures or the quality of life for social groups.

Here, we take the position that AI forecasts solely in the form of timelines (dates given by which we should expect to have developed transformative AI) are undesirable. To address this issue we propose a new AI forecasting framework along with a new scenario mapping technique that supports the framework. Independently, the framework and the new method each constitute novel contributions to the body of knowledge. However, together the framework and new technique demonstrate a holistic rethinking of how we forecast AI. It is this new perspective that we believe to be the paper's most significant contribution.

In the following pages the paper proceeds by first examining related literature. We do not consider the broader body of literature for the relevant topics, rather we focus only on the salient elements. After outlining scenario planning techniques, we move to propose a new subclass of scenario mapping techniques. Next, we propose a new method as part of this subclass which we call judgmental distillation mapping. This new method is then described as a critical component of the new AI forecasting framework. Following this description of the framework, we discuss strengths, weaknesses, the implications of practice and the implications on future research in AI forecasting. We conclude by summarizing the key ideas and recommendations.

2. Literature Review

This section examines several bodies of literature relevant to the holistic framework being proposed. This literature review is by no means comprehensive, and, due to the large number of academic disciplines and techniques covered, a more extensive literature is suggested for future work. We consider the research topics of forecasting, technology forecasting, scenario analysis, AI forecasting as well as a brief discussion of digital platforms.

2.1. Forecasting

Forecasting techniques are commonly broken down into two broad classes: judgmental methods and statistical methods [14]. Statistical methods are preferred for most forecasting applications and can range from simple extrapolations to complex neural network models or econometric systems of simultaneous equations [15]. However, statistical methods perform poorly in cases with little or no historical data, cases with a large degree of uncertainty and cases involving complex systems [16]. In such situations it is common to fall back on judgmental techniques. In this subsection we will forgo any discussion of statistical methods to focus on the different judgmental techniques and the challenges of expert selection.

Surveys are likely the most widely used judgmental technique. They solicit expert opinion from multiple experts without interaction between them. This technique is widely used because it is straightforward to implement and relatively inexpensive [1]. Challenges to this method include sampling difficulties, especially those due to nonresponses. The Cooke method (or the classic method) of assessing the quality of expert judgements for expert elicitation comes from the field of risk analysis [17]. It is a very powerful technique that involves the inclusion of calibration questions to calibrate the experts' forecasts so that they may be weighted during aggregation [18].

The Delphi technique was developed at the Rand Corporation in the 1950s at the same time as the development of scenario planning methods [19]. This approach involves a group of experts participating in an anonymized forecasting process through two or more rounds [1]. Each round involves answering questionnaires, aggregating the data and exchanging the summarized results and comments. Expert participation, expert selection and the falloff rate of participants over iterative survey rounds are the primary challenges. The Delphi technique is powerful and versatile, with the capability to be used for scenario building exercises as well as forecasting, and the flexibility to support large groups of experts with small modifications [20]. Despite its wide use for over a half century, there are still many questions about fundamental issues of its effectiveness for certain situations [21,22]. Specifically, academic work that has been conducted on the Delphi technique has frequently used students, which, for numerous reasons may be misleading. No work on the Delphi technique or other powerful judgmental forecasting techniques has been conducted to assess the quality of forecasts for the purpose of technology forecasting.

Prediction markets are exchange traded markets intended for predicting the outcomes of events. They rely on a platform that allows people to make trades depending on their assessment of these outcomes. Prediction market contracts are binary options that are created to represent a forecasting target and then traded through the market. During 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. The free market tends to collect and aggregate predictive information well due to the strong economic incentives for better information. Consequently, prediction markets often produce forecasts that have lower prediction error than conventional forecasting techniques [23]. Green et al. performed a comparison of the Delphi technique and prediction markets, finding that, when feasible, prediction markets have some advantages, but that the Delphi technique was still generally underused (it is unclear which is better for technology forecasting) [24]. However, the advantages of each technique were also dependent on the problem. Prediction markets performed better for short-term, straightforward problems whereas the Delphi technique was useful for a broader range of problems and for high uncertainty situations.

Superforecasting is a recently developed technique that uses 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 and to forecast certain types of targets (e.g., geopolitical events) better than any other methods [25]. The technique was developed using forecasting tournaments for a competition for the US' Intelligence Advanced Research Projects Activity (IARPA). The project was funded for the purpose of developing new methods in order to improve the US intelligence communities forecasting abilities [26]. However, 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 for predictions that are seemingly impossible. It is also not suitable for existential risk applications [27]. Furthermore, while it may be one of the most powerful forecasting methods available for near-term forecasts, it still is not able to make forecasts any better than a coin toss for events over five years in the future (Tetlock considers experts no better than normal people at forecasting political events).

Combining different types of forecasts that draw from different information sources can be a powerful technique for forecasting when there is significant uncertainty about the situation or uncertainty about the different methods [14]. Another powerful technique can be the adjustment of statistical forecasts using expert judgment, particularly in cases of high uncertainty where domain expertise is critical and environments are poorly defined [28]. In such cases, structured judgmental adjustment can be a very powerful technique as long as efforts are made to counter cognitive biases [29].

Scenario planning methods are sometimes considered an adjunct forecasting method and scenarios are commonly employed to deliver the results of forecasts to decision makers [14]. However, substantial work has been conducted considering their practical use in improving decision making under uncertainty [30,31]. They are considered an essential technique in the technology forecasting and management literature [1], thus, we devote an entire subsection to them in the following pages.

Goodwin and Wright examine both statistical and judgmental forecasting methods in their ability to aid the anticipation of rare, high-impact events [21]. They find that while all methods have limitations, it is possible to combine dialectical inquiry and components of devil's advocacy with the Delphi technique and scenario planning techniques to improve the anticipation of rare events. In their comparison of techniques (including an informative table comparing methods for anticpating rare events) they consider several judgmental methods including expert judgment, structured judgmental decomposition, structured analogies, judgmental adjustment and prediction markets, as well as the Delphi technique and scenario planning.

Selecting experts is a challenging but necessary task when using any of these judgmental forecasting techniques. The first step in identifying experts is to identify the range of perspectives that will be needed in the study [1]. Researchers typically want to prioritize the most knowledgeable experts for vital perspectives first; less vital viewpoints can often times use less knowledgeable experts or substitute secondary sources for expert opinion. Researchers should also be cognizant of possible sources of experts' biases when selecting experts and analyzing their responses. Some significant attributes include 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 [32]. 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, however, when it is necessary, documentation for the particular type of study commonly addresses these issues [33].

2.2. Technology Forecasting

Technology forecasting is a challenging task and the body of literature concerning this topic is very broad. To be certain, there is not a well-developed field of study that directly concerns the forecasting of future technologies. Much of what is considered here as technology forecasting literature is focused on technology management, and, consequently, many of the techniques are intended to aid in organizational management and planning.

A wide variety of methods are used for technology forecasting, including both statistical and judgmental techniques. Other techniques are also used, some of which are unique to technology forecasting. Innovation forecasting techniques can be used for mapping scientific domains that rely on bibliometric analysis [34]. Tech mining is a similar technique that harnesses data mining methods to extract information from patent databases and the Internet for the purposes of innovation forecasting [35]. Due to the substantial uncertainty, scenario analysis techniques are also widely used for strategic planning involving emerging technologies [1]. This subsection does not revisit judgmental forecasting techniques discussed in the previous subsection, but focuses on techniques that have not yet been discussed. Scenario analysis is discussed in depth in the following subsection.

Assessing progress—particularly the rate of progress—is essential when developing any type of technology forecasting model. This is so because the naïve assumption that historical trends can be extrapolated to the future is many times correct, and, consequently, trend extrapolation is a very powerful forecasting technique [1]. Indicators are variables that can be used for extrapolation or for building statistical forecasting models because we believe them to be good for predicting future progress. There are two basic types of indicators that are of interest in technology forecasting: science and technology indicators and social indicators. 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 indictors that suffice as best as 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.

Technology roadmapping is a widely used and flexible technique that is commonly used for strategic and long-term planning [36]. It is known to be particularly effective in structuring and streamlining the research and development process for organizations [37], but it can be used for planning at both an organizational level and a multi-organizational level. It is generally thought to consist of three distinct phases—a preliminary phase, a roadmap construction phase and a follow-up phase—and commonly uses workshops for the map generation phase [38]. When applied, it often uses a structured and graphical technique that enables exploring and communicating future scenarios. However, its lack of rigor and heavy reliance on visual aids can also be seen as weaknesses [1].

Innovation forecasting is a term that is typically associated with the use technology forecasting methods in combination with bibliometric analysis [34]. In general, bibliometric methods are powerful analysis tools for understanding the progression of science. Such methods have been used for the mapping of this progression in different scientific disciplines for several decades [39]. Maps of relational structures present in bibliometric data are useful for visualizing the state of research within the domain(s) of interest and can lead to insights regarding future research directions and geopolitical issues [40].

Tech mining is another notion that is frequently associated with innovation forecasting and management [35]. It generally refers to a broad set of techniques which can be used to generate indicators from data. Porter and Cunningham discuss the use of innovation indicators for understanding emerging technologies, and propose nearly 200 such indicators. Here, we consider tech forecasting to encompass all bibliometric and scientometric techniques used for the purposes of technology forecasting.

While we have focused here on judgmental forecasting techniques and other techniques for technology forecasting, there is evidence that suggests that extrapolation and statistical methods are better for forecasting technological progress [41]. Studies have found that technology forecasts developed using statistical methods were more accurate than those developed from other methods, with forecasts about autonomous systems and computers being the most predictable [42]. However, there is certainly not agreement on this topic. Brynjolfsson and Mitchell conclude that "simply extrapolating past trends will be misleading, and a new framework is needed," [43]. The holistic perspective proposed here attempts to provide a new framework.

2.3. Scenario Analysis

Scenario analysis is a term used in technology management literature to refer to scenario planning techniques when applied in the context of technology and innovation forecasting [1]. People use scenario analysis naturally by thinking in terms of future scenarios when making most decisions involving uncertainty in everyday life. It is also a very effective technique for decision-making processes in more complex situations [44]. Scenario methods are rooted in strategic planning exercises from the military in the form of 'war game' simulations, or simply wargames. Wargames are a type of strategy game that have both amateur and professional uses. For amateurs they are used for entertainment, with some of the earliest examples being the games of Go and chess. Fantasy role-play games such as

Dungeons and Dragons are also derived from wargames and used for entertainment. Professionally, wargames can be used as a training exercise or for research into plausible scenarios for highly uncertain environments such as those encountered on battlefields during wartime [45]. Events in World War II, such as the allied preparations for D-Day, made clear to military commanders the value of wargames and scenario techniques. Following the war, during the 1950s and 1960s, new scenario techniques were independently developed in both the United States and France. In the United States, the methods were developed at the Rand Corporation, a research and development venture of the US Air Force. In France, the techniques were developed for public planning purposes. Although developed independently, these two schools eventually led to the development of very similar scenario techniques.

Scenario analysis, as it is known today, typically involves the development of several different scenarios of plausible futures. It is most widely thought of as a qualitative technique for the purposes of strategic planning in organizations [46]. 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 many quantitative and hybrid techniques have also been shown to be practically useful [20]. Here we describe three schools of scenario techniques: the intuitive logics school, the probabilistic modified trends (PMT) school and La Prospective, a.k.a. the French school. We attempt to outline these different schools below.

The most prominent of qualitative methods, having received the most attention in the scenario planning literature, is the intuitive logics school [20]. After being developed by at the Rand Corporation in the 1950s and 1960s, it was popularized from its use by Royal Dutch Shell in the 1970s, and it is sometimes referred to as the 'Shell approach' [19]. 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 [47].

The most common of quantitative methods is considered to be the PMT school, which also originated at the Rand Corporation during the 1960s [20,48]. This school incorporates two distinct methodologies: trend-impact analysis (TIA) and cross-impact analysis (CIA) [19]. 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 during the cross-impact calculation.

The two schools described above may do well to illustrate qualitative and quantitative scenario techniques, but they 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 notably 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 [49]. 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 [20]. 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 several computer-aided tools [19]. Although it arose independently, this school can also be seen to a large extent to combine the intuitive logics and PMT methodologies. A full review of scenario planning literature is beyond the scope of this work, but we believe that these simple characterizations to be sufficient for the purpose of this work.

2.3.1. Scenario Analysis for Mapping

Traditional qualitative scenario planning techniques certainly have a role in assisting decision makers of organizations and other stakeholders involved in the development and governance of transformative AI. However, such techniques can do little to map the plausible paths of AI technology development due to the large space of possible paths. Traditional quantitative methods certainly have a role in some organizational decisions as well. However, while they are commonly sufficient for strategic decision making, they typically fall short for understanding and informing design decisions of complex systems.

Over the past two decades, the use of scenario analysis techniques for mapping complex systems, complex environments and complex technologies has increased [50]. Particularly, we focus on three such techniques. The first is a relatively obscure method that has seen little practical application, yet it has significant potential for mapping the paths of possible futures for which there are high levels of uncertainty [20]. The second originated as a way to represent social scientific knowledge through directed graphs, and has since become a common method for scenario analysis in multi-organizational contexts [51]. 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. Each of these techniques offers insight that contributes to the holistic framework proposed here for forecasting transformative AI.

Scenario network mapping (SNM) is a qualitative scenario technique that was proposed to improve upon existing methods by including a substantially larger number of scenarios, each of which forms a portion of a particular pathway of possible events [52]. This results in a network-like structure which is easily updated in the future with the addition, removal and repositioning of scenarios and their interactions in light of new information. A key feature to SNMs is their reliance on the holonic principle which implies that a scenario can also be decomposed into more scenarios. Following the development of the scenario map, the scenarios can be refined further using causal layered analysis techniques [53]. This technique benefits from larger groups of experts, because the structure of the network becomes more comprehensive with iterative refinement. In a typical SNM scenario building workshop, several hundred possible scenarios are generated, which are then typically reduced to 30–50 plausible scenarios that are used to create the scenario map. Due to this ability to accommodate a large number of plausible scenarios, we see potential for this method for its intended purpose as well as the potential for some derivative of it to be effectively used to identify a large number of possible paths to HLAI (here, we denote the later SNM with an asterisk).

Axelrod first introduced cognitive maps in the 1970s to represent social scientific knowledge with directed graphs [54]. His work has since been extended to a variety of applications including scenario analysis. However, the psychological notion of cognitive maps—people's representations of their environments and mental modes—comes from two decades earlier [55]. Cognitive maps are effective for facilitating information structuring, elaboration, sequencing and interaction among participants or stakeholders [51]. They are sometimes thought of as causal maps because of the causal network of relationships represented in the nodes and edges. Here nodes can be thought of as scenarios, and the edges describe the causal relationships between them.

Fuzzy cognitive map (FCM) modeling is another hybrid scenario technique that can better integrate expert, stakeholder and historical data through the development of scenarios that assist in linking quantitative models with qualitative storylines [50,56]. FCMs were first proposed by Kosko in the 1980s as a means for making qualitative cognitive maps—used for representing social scientific knowledge [54]—computable by incorporating fuzzy logic. While effective for scenario analysis, FCMs are used generally for decision making and modeling complex systems, and they have a wide variety of applications in multiple domains ranging from online privacy management to robotics [57]. Simply, we can think about FCMs as weighted directed graphs wherein the nodes are fuzzy (i.e., they take a continuous value from zero to one rather than a discrete value) and representative of verbally described concepts while the edges are representative of causal effects.

2.3.2. Using Expert Opinion for Scenario Analysis

Virtually all scenario analysis 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 [20]. 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 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. Based on six case studies, List suggests that for creating SNMs four half-day workshops with 20 experts is roughly optimal [58]. 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 expert's judgement [56].

2.4. AI Forecasting

The study of forecasting AI and HLAI is in its nascency, and 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 [59]. Since 2006 12 more surveys have been administered [60]. Such surveys have been used to generate forecasts in the form of timelines. The most recent work has aggregated probability distributions collected from participants [10,61,62]. 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 forecasts from surveys of expert opinion, the foremost perhaps being the questionable reliability of experts [25].

The most rigorous of expert survey studies include four particular surveys which have been conducted since 2009, all pertaining to notions of artificial general intelligence (AGI). (Here, we consider human-level machine intelligence, high-level machine intelligence, human-level artificial intelligence and other similar ideas as notions of AGI.) The first of these surveys was conducted at the 2nd Conference on Artificial General Intelligence and found that the majority of experts believed that HLAI would be realized around the middle of the 21st century or sooner [63]. The 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. 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 [64]. This survey questioned participants as to when they expected high-level machine intelligence (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. Slightly over half of them believed that this would be a positive development while roughly 30% expected it to have negative consequences.

The next survey solicited the primary authors of the 2015 Neural Information Processing Systems (NeurIPS) conference and the 2015 International Conference on Machine Learning (ICML) [61]. This study questioned participants on their forecasts of HLMI, but also included questions about a large number of specific tasks. All forecasters were asked for 10%, 50% and 90% probabilities, which effectively elicited a probability distribution from each. This was not new, 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, but, interestingly, a median of 120 years before all human jobs would be automated. The study also found Asian participants to have much earlier predictions that Europeans and North Americans.

The most recent expert survey was solicited at the 2018 International Conference on Machine Learning, the 2018 International Joint Conference on Artificial Intelligence and the 2018 Joint Conference on Human-Level Artificial Intelligence [10]. Rather than focusing on notions of AGI, this study elicited five forecasts for different levels of transformative AI. It also included calibration questions—the first expert survey in the context of AI forecasting to do so. While the forecasts were closely aligned with the previous study, an improved statistical model was used. The use of a naïve calibration technique improved the explainability of the variability in the statistical model for the most extreme transformative AI forecasts. The results also indicated that forecasts from researchers at the HLAI conference were more precise and that this group exhibited lower levels of uncertainty about their forecasts.

A number of meta-analyses of AI forecasting studies have also been conducted. In 2012 and 2014, Armstrong and Sotala and Armstrong et al. assessed previous timeline predictions that had been incorrect [65,66]. They proposed a decomposition schema for analyzing, judging and improving the previous predictions. Muehlhauser has also conducted examinations of timelines and previous AI forecasts [67,68]. His studies offer the most comprehensive discussion of timelines for notions of AGI prior to the surveys conducted over the past decade. Regarding timelines, Muehlhauser concludes that we have learned very little from previous timelines other than the suggestion that it is likely we achieve AGI sometime in the 21st century. He further explores what we can learn from previous timelines and concludes with a list of ten suggestions for further exploration of the existing literature.

AI Impacts (www.aiimpacts.org) is a non-profit organization that is commonly thought to be the leading AI forecasting organization. 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 aimed at understanding the potential impacts and nature of HLAI. AI Impacts has contributed significantly to practical forecasting knowledge, even leading a major AI forecasting study in 2016 [61].

Recent work by Amodei and Hernandez presented a trendline for the increase in training costs for major milestones in AI progress between 2012 and 2018 [69]. This trendline depicted exponential growth for the increase in the amount of training time required for achieving selected AI milestones; the training time doubled every 3.5 months. However, several critiques of this have emerged [70,71], the most compelling being that from a purely economic perspective the trend was unsustainable for a long period; the exponential rate for training costs was significantly greater than the exponential decrease in costs of compute. Despite these fundamental challenges to the trend, AI experts generally expect the trend to continue for at least 10 years [10]. While not receiving as much visibility, other efforts have been made to plot or collect relevant data to measure the progress of AI research [72–75]. Despite these efforts, the best technology indicator given the criteria previously discussed may be that of Amodei and Hernandez.

While most practical work on AI forecasting to date has relied on expert surveys and extrapolation, there are several important exceptions. Zhang and Dafoe recently conducted a large-scale survey of non-expert opinion that was intended to assess the opinions of the American public regarding AI progress [62]. Another study that was conducted in 2009 used the technology roadmapping technique in an attempt to create a roadmap for the development of HLAI [76]. The results of this workshop depicted expected milestones on the path to HLAI arranged in a two-dimensional grid of individual capability and sociocultural engagement. While organizers of the workshop were disappointed in what they perceived as a failure of the workshop to generate a straightforward roadmap [77,78], arguably 50% or greater of the tasks have been completed [79]. More recently, the Association for the Advancement of Artificial Intelligence and the Computing Community Consortium have completed a 20-year Roadmap for AI research [80]. This roadmap was less ambitious than the earlier attempt, and focused on three major themes: integrated intelligence, meaningful interaction and self-aware learning. The AI Roadmap Institute (www.roadmapinstitute.org) has also been created to study, create and compare roadmaps to AGI. Although the institute's efforts have resulted in the development of a roadmap, however, it does not concern technical elements as much as social elements.

Another relevant body of research concerns risk analysis. Significant work has been conducted regarding existential risks, and, particularly, the risks posed by AI and superintelligence. (For the purposes of forecasting we do not consider superintelligence, an intelligence explosion or their ramifications [81]). In 2017 Baum created a comprehensive survey of AGI projects that includes a mapping of all relevant stakeholders at the time [82]. Barrett and Baum conducted two studies during 2017, one which focused on the use of fault trees and influence diagrams for risks analysis, and the second which considered expert elicitation methods and aggregation techniques as well as event trees, including a probabilistic framework [83,84]. Additionally, in 2017, Baum et al. examined techniques for modeling and interpreting expert disagreement about superintelligence [85].

Recent work evaluated the methods currently used to quantify existential risk, considering both statistical and judgmental forecasting methods [27]. This study found that while there were no clear 'winners,' the adaptation of large-scale models, the Delphi technique and individual subjective opinions (when elicited through a methodologically rigorous process) had the highest potential. Furthermore, the authors concluded that surveys met all of the criteria to an acceptable degree and that fault tress, Bayesian networks and aggregated expert opinion were all well suited for quantifying AI existential risks. Prediction markets and superforecasting were not found to be especially suitable in general, or for AI risks specifically.

Other recent work by Avin has considered AI forecasting from the futures studies perspective including consideration of the use of scenario planning and wargaming as well as standard judgmental and statistical methods [86]. Wargames (a.k.a. professional role-playing games or government simulation games) are particularly promising as they can be used for informing difficult and complex strategic policy decisions [87]. As mentioned earlier, wargaming can serve two valuable purposes in preparing organizations for futures involving a large degree of uncertainty—training and research—and has been suggested by Avin (in the form of an AI scenario role-play game) as a valuable tool for both of these purposes in the AI strategy field. Furthermore, wargaming can be used in a model-game-model analysis framework to iteratively refine different models for how certain future scenarios may unfold [88]. The work of Beard et al., Barrett and Baum, and Avin represent the only known work in the literature to explore the possibilities for judgmental forecasting techniques (other than surveys) and scenario planning techniques for AI strategy purposes in any depth [27,83]. (We note that there are ongoing efforts to use and improve prediction markets for AI forecasting as well as to develop a new type of forecasting platform for AI forecasting.)

Assessing progress in AI is crucial in order to use extrapolation or other statistical forecasting techniques that require historical data. Consequently, a substantial amount of work has considered theoretical aspects of assessing and modeling AI progress for different ends [89–91]. This discussion focuses on some recent efforts and other notable contributions. In 2018, Martinez-Plumed et al. proposed a framework for assessing AI advances using a Pareto surface which attempted to account for neglected dimensions of AI progress [92]. More recently, in 2019, Martinez-Plumed and Hernandez-Orallo built on previous work on item response theory to propose four indicators for evaluating results from AI benchmarks: two for the milestone or benchmark; difficulty and discrimination, and two for the AI agent; ability and generality [93]. Hernandez-Orallo has written extensively about measures of intelligence intended to be useful for intelligences of all substrates that allow for our existing anthropocentric psychometric tests to be replaced with a universal framework [94]. In contrast to these studies, a measure of intelligence has been proposed by Riedl that is based on creativity [95]. The topic has also garnered mainstream attention with a workshop being dedicated to it in 2015 [96], and work on it being featured in Nature in 2016 [97]. Although there is no consensus on how to measure AI progress or intelligence, it is clear that simple measures which can be represented in a small number of dimensions are elusive.

Work by Brundage has attempted to develop a more rigorous framework for modeling progress in AI [6]. In it he suggests that this type of rigorous modeling process as being a necessary precursor to the development of plausible future scenarios for aiding in strategic decision making. To these ends he proposes an AI progress modeling framework that considers the rate of progress in hardware, software, human input elements and specialized AI system performance. In this work, Brundage considered indicators and statistical forecasts as being fundamental for modeling AI progress. However, later efforts by Brundage began to try and integrate scenario planning and judgmental forecasting techniques into formal models (i.e., agent-based models and game theory) [98]. While this later work did not result in the proposal of a rigorous framework like his earlier effort, we believe that it indicates that the integration of various techniques is necessary for adequately modeling and forecasting AI progress. Moreover, it was successful in identifying numerous challenges posed by such an integration. The proposed framework here draws from this previous work in attempts to address these challenges.

In the AI governance research agenda, Dafoe discusses the notion of mapping technical possibilities, or the technical landscape, as a research cluster for understanding possible transformative futures [7]. He also notes the important role of assessing AI progress and modeling AI progress. Separately, he discusses AI forecasting and its challenges and also includes a desiderata for forecasting targets. This paper addresses the issues of generating a mapping, and the task of forecasting events comprising this mapping to the greatest degree that we are able to, as the AI governance research agenda prescribes.

Table 1 compares existing studies to illustrate to readers the focus on surveys and the lack of focus on alternative techniques (it also demonstrates the little amount of work existing in the literature). The earlier work of Brundage seen in Table 1 and discussed in the previous paragraph is the only other known work to consider an integrated and rigorous methodical approach to the specific problem of AI forecasting. However, Brundage did not consider any of these techniques from a forecasting perspective in the way we do. Particularly, his work focuses on applying these techniques directly to a model of AI governance. This study goes further by building on the need for a mapping described by Dafoe and by considering a broader range of forecasting and scenario analysis techniques than previous work to develop a holistic forecasting framework.

Table 1. A comparison of surveys and alternative studies previously conducted on AI forecasting.

Study	Year	Type	Results	Conclusion (Median yrs)
		AI Forecasting Surveys		
Baum et al.	2011	Expert (HLAI)	Statistical	Experts expect HLAI in coming decades, much disagreement
Grace et al.	2016	Expert	Probabilistic	45 yrs 50% chance HLAI, Significant cognitive dissonance
Gruetzemacher et al.	2019	Expert (HLAI/AI)	Probabilistic	50 yrs 50% chance HLAI, Type of expertise is significant
Müller and Bostrom	2014	Expert	Statistical	2040–50 50% chance HLAI; <30 yrs to superintelligence
Zhang & Dafoe	2019	Non-expert (Americans)	Probabilistic	54% chance of HLAI by 2028, support AI, weak support HLAI
		Other AI Forecasting Studies		
Amodei and Hernandez	2018	Extrapolation	Trendline	Compute required for AI milestones doubling every 18 months
Armstrong and Sotala	2012	Comparative Analysis	Decomposition schema analysis	Expert predictions contradictory and no better than non-experts
Armstrong et al.	2014	Comparative Analysis	Decomposition schema analysis	Models superior to judgment, expert judgment poor, timelines unreliable
Brundage	2016	Methods	Modeling Framework	A framework for modeling AI progress

Table 1. *Cont.*

Study	Year	Type	Results	Conclusion (Median yrs)
Muehlhauser	2015	Comparative Analysis	Generalization	We know very little about timelines, accuracy is difficult
Muehlhauser	2016	Historical Survey	Suggestions	Future work ideas, AI characterized by periods of hype/pessimism

2.5. Summary of the Related Literature

There are generally thought to be two types of forecasting techniques: judgmental and statistical. Statistical methods are typically preferred when data is available; however, in cases for which data does not exist, is missing or for which there are other inherent irreducible uncertainties, judgmental techniques are commonly the best or only options. AI forecasting falls into the later of these categories. Work from Brundage has previously proposed a general framework for modeling AI progress [6], and later work attempted to integrate scenario analysis, expert judgment and formal modeling [98]. Although most previous studies using judgmental techniques have used expert surveys, there is new evidence that other techniques are more appropriate for this problem [27]. Other potentially valuable techniques, such as tech mining, bibliometric analysis or mapping the technical possibilities have been suggested but have not been attempted in the literature. This study goes further than previous work by considering a holistic framework which attempts to use statistical techniques as best as possible, and to augment their use by including judgmental techniques and scenario analysis techniques. We ultimately take a step beyond forecasting to suggest exercises for strategic management and planning.

3. Judgmental Distillation Mapping

Section 2.3.1 highlighted three scenario analysis techniques that have mapping qualities. We refer to these techniques collectively as scenario mapping techniques due to two significant properties they share: (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 (i.e. maps). Although only three have been identified, other approaches are possible. Here, we draw from the existing techniques to propose a new scenario mapping technique which also exhibits the same mapping characteristics as the techniques described in Section 2.3.1. We refer to the proposed technique as judgmental distillation mapping (JDM). Figure 1 depicts a diagram of the JDM process.

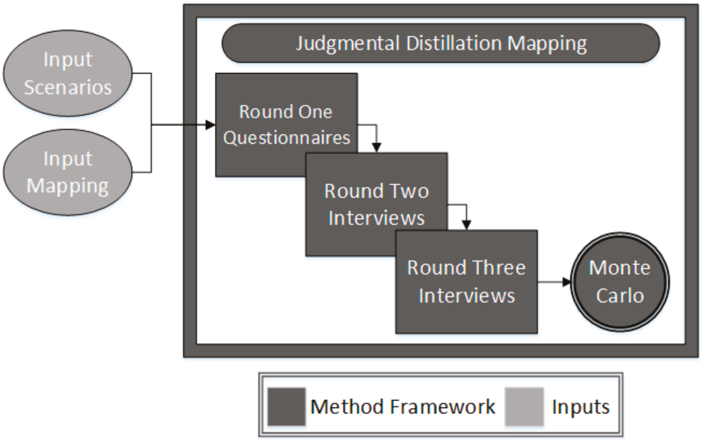
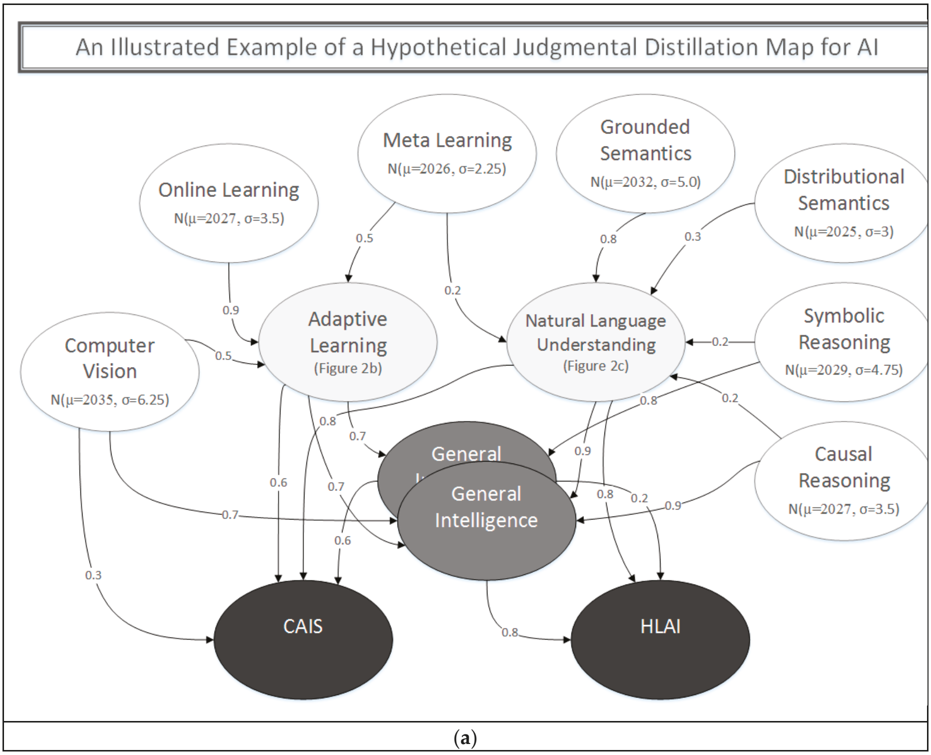


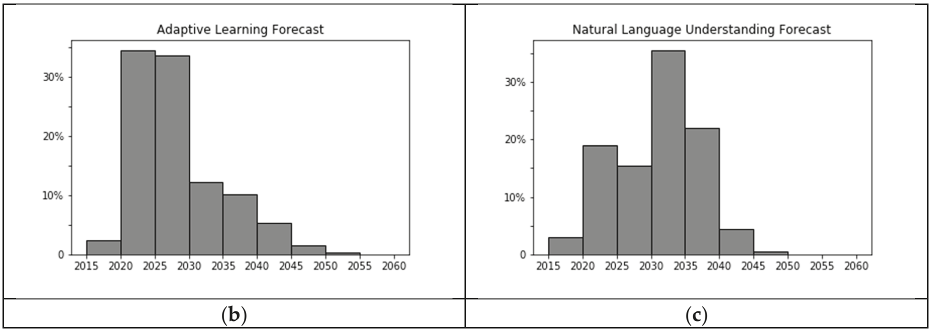
Figure 1. The judgmental distillation mapping technique. The technique is flexible and can be thought of as generally being comprised of iterative rounds of questionnaires and interviews intended to isolate a scenario map for which forecasts are generated through Monte Carlo simulation.

The map created (see Figure 2) is distilled from larger input maps (we use map here generically to refer to the combination of tech mining and historical indicators) comprised of both historical or tech mining data, and scenarios developed either by previous rounds of the judgmental distillation process, through interviews or through a technical scenario network mapping workshop (i.e., a scenario network mapping solely for mapping paths to notions of AGI [99]). The scenario map (i.e., the graph) is equivalent in characteristics to that of an FCM, with advanced technologies being represented as nodes. The input nodes represent technologies for which forecasters believe it tractable to use existing forecasting techniques to forecast. The 2nd order and greater nodes in the maps cannot be forecast directly using powerful, traditional techniques such as Delphi, prediction markets or superforecasting. However, these methods are suitable for the first order nodes as long as they are used in a fashion that generates the probability distributions that are necessary for computing the timelines for higher order technologies. These timelines are generated using Monte Carlo simulation and the causal relations between technologies, as determined by expert judgment given the input data. As the final outcomes of transformative AI technology are unknown (possible outcomes include HLAI, comprehensive AI services or AI generating algorithms [100]), the resulting map is able to accommodate a variety of outcomes. Figure 2 depicts a possible result of the JDM process. (This figure is not intended as a forecast, but rather as an example of what JDM could result in. Input distributions are randomly assigned using a normal distribution as opposed to aggregated, adjusted results from JDM.)



(a)

Figure 2. *Cont.*



(**b**) (**c**)

Figure 2. (**a**) A hypothetical judgmental distillation map is depicted. White ovals are inputs and light grey ovals are next generation (2nd order) technologies. General intelligence is depicted in a stacked fashion to indicate the possibility of future technological scenarios in the model to be realized through the combination of different paths (i.e., adheres to the holonic principle). The links in the figure are representative of causal relationships and the weights for these links correspond to the strength of these relationships. Note that this figure is not intended to be a forecast, but rather an example of what the JDM process could result in. Input distributions are randomly assigned using a normal distribution. Actual input distributions would not be based on a normal distribution and would be aggregated from expert opinion rather than parameterized distributions. (**b**) A histogram depicting the results of a Monte Carlo simulation for the next generation adaptive learning technology. (**c**) A histogram depicting the results of a Monte Carlo simulation for the next generation natural language understanding (NLU) technology. (Monte Carlo simulation is used to generate the distributions found in b and c. A notebook for computing these distributions can be found here: www.github.com/rossgritz/research/.)

JDM is a resource intensive technique that requires a substantial degree of expertise from the forecaster(s) as well as a large number of participating experts. The primary burdens of decomposition and aggregation fall to the facilitator of the process, and, as noted, this provides significant opportunity for facilitators to exercise their own judgment. If not making all input data available to all experts, substantial effort would be required in delegating portions of the input data, and subsequently developing individualized interview questions and questionnaires for the participating experts. To obtain the best results, it is likely best to employ a team for the entire JDM process. Moreover, the best results from the process may be achieved with long periods dedicated to judgmental distillation.

There are three primary inputs for the JDM process: historical data and statistical forecasts, judgmental data and forecasts, and scenarios. Historical data, statistical forecasts and scenarios should initially be in the form of mappings, but such inputs will frequently be deconstructed by the forecasters into forms easily digestible for analysis. Efforts to avoid information overload should be prioritized by forecasters when presenting the information to experts. The questionnaire and interview questions involve asking experts to respond to or comment on scenarios, statistical forecasts, judgmental forecasts and the relationships between these input items. While an example of data that could be shown to interview candidates is shown in Figure 3 and described below, the majority of the JDM process may still be comprised of questions building on qualitative input data rather than quantitative input data.

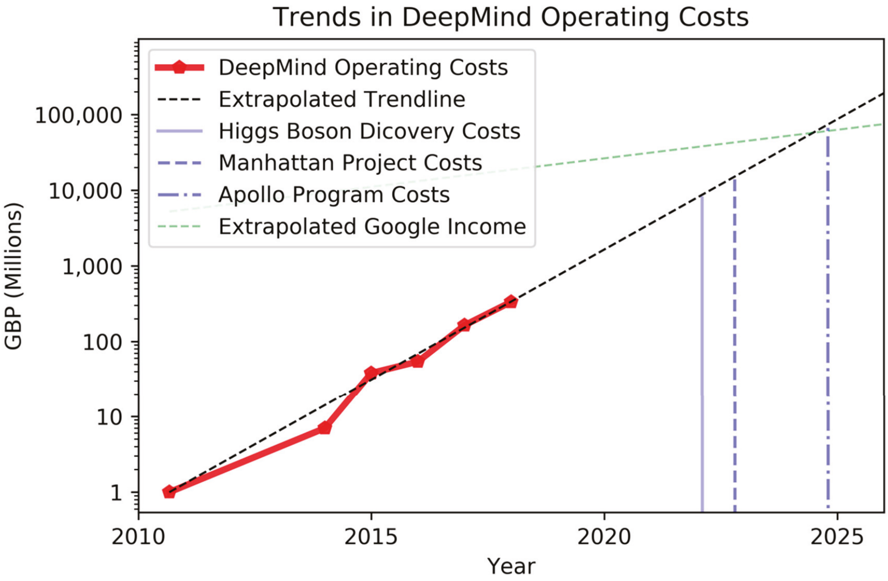


Figure 3. This depicts a simple extrapolative forecast of a social indicator. This is an example of the type of quantitative information that can be provided to experts for adjustment, distillation and aggregation. When presented with this, experts could be asked whether they agree or disagree that this extrapolation is reasonable. If they disagree, they would be asked to explain how they disagreed and what they thought was a reasonable trend for the indicator presented in the figure. Based on these responses, they may also be questioned about economic realities governing the behavior of this indicator and whether they believed it was possible, even over a substantially longer timeframe, for these economic factors to be altered such that this the indicator may ultimately hit some of the major milestones depicted. They may also be asked questions raising concerns identified by other experts or questions as to why or why not AI research should be analogous to nuclear physics or rocket science. Careful consideration about the indicators and the questions to ask would be determined by the forecaster, or by a forecasting team.

Table 2 is included below to enable an easy comparison of the scenario mapping techniques for readers. It is inserted here so that the newly proposed method of JDM can be included. Therefore, it demonstrates how the new method just described is the only scenario mapping method that is capable of producing probabilistic forecasts for a large number of complex scenarios. The table also indicates that this increased value does come at the cost of substantial resources and a large number of experts (this is discussed further in the discussion section). These factors increase the practical applicability of the method significantly. It can also be noted here that SNM is another useful method for complex cases with large numbers of possible scenarios. Specifically, SNM is useful for mapping the paths to AGI qualitatively while JDM is better suited for generating probabilistic forecasts for AGI and other transformative AI technologies. Overall, (considering the AGI-SNM workshopping technique under development) this table clearly demonstrated the significant practical advantages of the newly proposed methods over the other methods grouped into the scenario mapping class of techniques. (While these new techniques are clearly better for the purposes of forecasting AGI and transformative AI, they may also be useful for other forecasting applications, e.g., for forecasting issues related to existential risks, for forecasting issues related to other complex technology development or for forecasting the broader progress of different domains of scientific study. We do not discuss these options here, but we do suggest that future work consider this broader range of alternate applications.)

Table 2. A comparison of scenario mapping techniques.

Scenario Mapping Techniques					
Technique	**No. Scenarios**	**Quantitative**	**Qualitative**	**Strengths**	**Weaknesses**
Scenario Network Mapping (SNM)	30 to 50	No	Yes	Complex cases with large numbers of scenarios	Time consuming and requires 15–20 experts
Cognitive Maps	8 to 24	No	Yes	Useful in multiplle organization contexts, rapid workshop development	Weak scenario development, lack of rigor in method
Fuzzy Cognitive Map (FCM)	8 to 18	Yes	Yes	Flexible method, quantitative and qualitative elements, aggregates coginitive maps	Limited quantitiative value, limited judgmental value
Judgmental Distillation Mapping (JDM)	6 to 30	Yes	Yes	Complex cases that require probabilistic forecasts	Resource intensive and requires diversty of experts

4. A Holistic Framework for Forecasting AI

JDM was developed to integrate a large variety of forecasts into a single mapping that includes probabilistic timelines for its components. However, these results are more model than narrative and they focus on technological developments rather than economic, political, social or resource-related factors. Moreover, JDM is a technique rather than a broader solution for forecasting, planning and decision making on a continuing basis. The proposed holistic framework uses JDM to leverage flexible combinations of powerful forecasting techniques in an attempt to provide a comprehensive solution.

The framework is depicted in Figure 4. In this figure, inputs are depicted as rectangles, required inputs are depicted as ovals and actionable forecasts are depicted as circles. As depicted in the legend, the elements of the framework can be thought to comprise three distinct groups of processes: input forecasts, JDM and strategic planning. The inputs are comprised of traditional forecasting and scenario analysis/mapping techniques. JDM is modified for the framework to generate two outputs, one directed back at the next iteration of inputs and the other directed at the strategic planning processes. These strategic planning processes then build on JDM forecasts by considering economic, political and technological aspects with both traditional scenario analysis techniques as well as a powerful existing scenario mapping technique. A strategic AI scenario role-playing game can be used for training as well as scenario refinement.

The inputs to the framework are illustrated in Figure 4, and are consistent with the input requirements described for JDM. The quality of the inputs is expected to be strongly correlated with the number of experts, the time requirements and the input forecast quality of JDM. Therefore, we anticipate tech mining, indicators (tech and social), interviews and survey results to all be relatively essential for obtaining a reasonable output given a reasonable amount of resources. The non-essential inputs depicted are SNM, the Delphi technique, superforecasting and prediction markets. The SNM input (scenario network mapping*) is not equivalent to the actionable SNM, but is a workshop technique focused on mapping the paths to strong AI [99] that uses a highly modified form of SNM specifically developed for this purpose. Alternately, the form of SNM used for strategic planning is consistent with the original intentions of the technique and is discussed at length below with the discussion of the strategic planning elements of the framework.

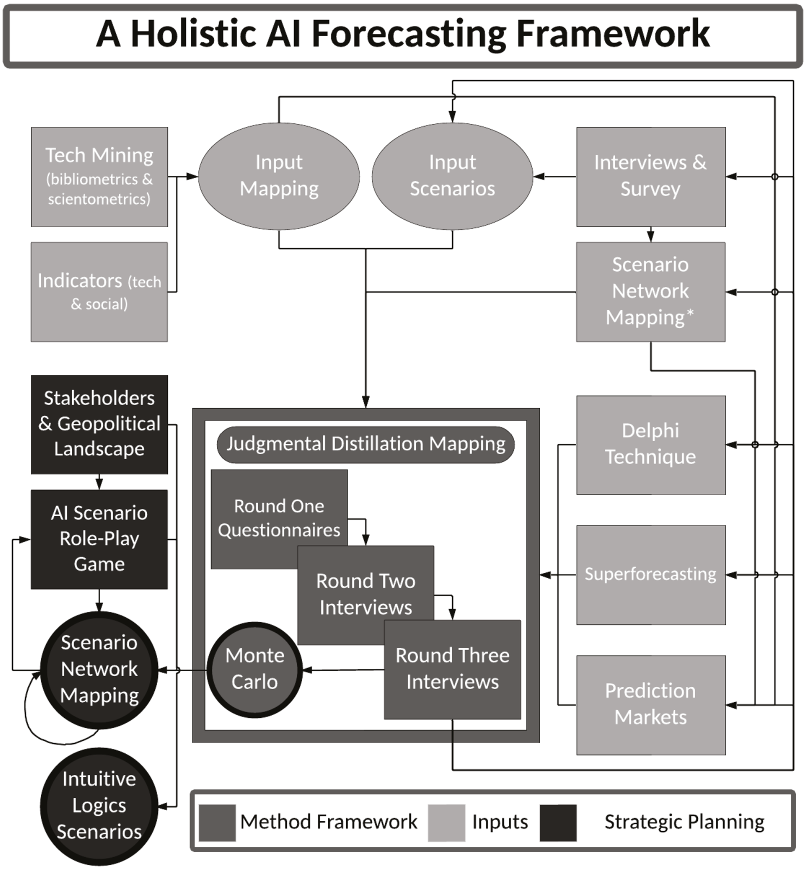

A Holistic AI Forecasting Framework
Tech Mining (bibliometrics & scientometrics)
Input Mapping
Input Scenarios
Interviews & Survey
Indicators (tech & social)
Scenario Network Mapping*
Stakeholders & Geopolitical Landscape
Judgmental Distillation Mapping
Delphi Technique
Round One Questionnaires
AI Scenario Role-Play Game
Round Two Interviews
Superforecasting
Scenario Network Mapping
Monte Carlo
Round Three Interviews
Prediction Markets
Intuitive Logics Scenarios
Method Framework
Inputs
Strategic Planning

Figure 4. The proposed holistic framework for AI forecasting. Rectangular boxes denote inputs, ovals denote required inputs and circles denote actionable forecasts. Inputs to the framework must include scenarios and a mapping of indicators, however, the specific choice of these and the methods for obtaining them are flexible.

JDM was described thoroughly in the previous subsection. However, in this discussion we considered only the output of the mapping and its timelines. When incorporated into the framework, JDM assumes the dual role of both generating forecasts and informing future component forecasts for iterative application of JDM in the holistic framework. In this way, the framework represents an ongoing forecasting process in which the judgmental distillation process has two objectives and two outputs; one for planning and one for continuing the forecasting process. The figure depicts the first output as moving left for informing actionable strategic planning techniques, and the second element as directing qualitative information right to be merged with updated indicators for developing new targets, forecasts and the next iteration of JDM. This second role performed by JDM, of providing feedback for future iterations, also has the effect of refining the previous forecasts and forecast targets. Since the process of completing an iteration of JDM in the proposed framework is resource intensive, it may be realistic to iterate over longer time periods, e.g., on an annual basis. This may be more amenable to expert participation because the frequency would be less burdensome to the experts, and with a relatively modest incentive (e.g., a gift card or lodging reimbursement for adding a workshop to an annual conference itinerary) participation may pose less of a challenge than other elements of the framework.

The input forecasts and the JDM process can be thought to work in tandem to produce an updated mapping with timelines on a continuing basis. It is likely that this cyclic pair of processes is sufficient to satisfy the requirements of the AI strategy community for a mapping and timelines [7]. However, we do not have to stop here. The purpose of forecasting is to inform decision makers such that they can make the best decisions, and, in order to do this we can draw from the methods discussed in the literature review to extend the JDM results so that they are most effectively used. The weakness of the mapping and timelines resulting from JDM is their heavy focus on the technology. To make the best strategic decisions, numerous factors must be considered (e.g., economic, political and technological). As discussed earlier, scenario analysis techniques are used to incorporate such factors in the planning and decision-making processes. Despite being strongly influenced by scenario analysis techniques, JDM does not include consideration of political factors, social factors or resource-related factors. Therefore, the strategic planning portion of the framework builds on the mapping and timelines for future AI technologies produced from JDM by considering economic, political, social and resource-related factors. It does this by two methods; one for high-level planning (intuitive logics scenarios) and another for exploring granular scenarios (scenario network mapping scenarios). No AI experts are required for these strategic planning elements in the framework. While we only consider the use of scenario planning to improve upon the results of the technology maps and timelines generated from JDM, it is also possible to extend the framework to incorporate forecasts of economic and political events into a separate JDM process keeping the technology map fixed. This possible dual use of JDM underscores the power of the holistic forecasting framework.

The use of intuitive logics scenario planning is inspired by their extensive record of success in business applications. Due to their widespread use and popularity, such scenarios may be more acceptable for use by traditional policy professionals not familiar with AI strategy or advanced scenario planning and technology forecasting methodology (their reliance on narratives makes them more palatable for such persons). In this case, the intuitive logics scenario planning technique would be used as intended to address and plan for uncertainties that are not implicit in the forecasts. This will likely lead to three or four high-level scenarios which can be used for guiding planning and decision-making processes directly. Scenario network maps may have some advantages over intuitive logics scenarios, however, they each can play important roles. Intuitive logics scenarios may be suitable for public dissemination whereas scenario network maps may be too granular and include sensitive information that may not be suitable for public release. Because intuitive logics scenarios may be more appropriate for politicians or other stakeholders who are not familiar with more advanced scenario planning techniques, they are sufficient for official reports from institutions using any holistic forecasting framework. When such scenarios are substantiated by a rigorous forecasting methodology, as discussed here, they can be much more effective for affecting public policy decisions.

SNM is also used here, i.e., the strategic planning context, in the manner that it was intended. However, the workshop technique will need to be modified from that detailed in the SNM manual [101] in order to incorporate the technology map generated from JDM. One of the unique advantages of SNM in this application is its use of the holonic principle. The holonic principle enables the deconstruction of complex scenarios even further. Therefore, Figure 4 includes a self-referential process arrow to indicate the possibility of continuing the scenario decomposition process to the desired level. This could be very useful due to the complex nature of this unique wicked problem [102].

JDM is also a flexible method, and within the holistic framework it can be used to forecast a large variety of AI forecasting targets. The output map would typically be expected to be similar to an FCM, comprised of somewhere between six and twenty nodes. However, there is no fundamental limit on the number of nodes in the map and it could be possible to retain the holonic principle from an input SNM through the distillation process to the output (or to use the holonic principle during the distillation process). Therefore, JDM could be used to forecast automatability of classes of jobs and even particular jobs or tasks with in specific jobs by means of judgmental decomposition. The details of such a process are not discussed at length here, but it is important to realize that as a forecasting project proceeds to further decompose its targets the required resources would continue to increase. Therefore, while it may be possible to use the framework for forecasting the automation of individual tasks, it is likely not a reasonable pursuit for most organizations due to resource constraints.

5. Discussion

5.1. Strengths and Weaknesses

One of the most obvious weaknesses of both JDM and the proposed framework are the heavy reliance of each on the use and elicitation of expert opinion. These methods may prove difficult to apply when access to experts is limited or biased (as in a single organization). Moreover, the resource requirements may be quite costly in the need for forecasting expertise. The JDM process requires a substantial amount of analysis on the part of the forecaster(s) for deconstructing, creating individualized expert questions and aggregating expert opinion into a single scenario map of AI technologies. The process, as envisioned here, is most likely better suited for teams when working on projects of any reasonable scale. However, the method proposed is flexible and could be revised so as to maintain the holistic framework while reducing reliance on expert judgment. It is unclear whether this would be desirable or not, but it is worth further examination.

The reliance of the proposed method and framework on expert judgment is also one of their greatest strengths. The literature review indicated that for forecasting problems concerning large degrees of uncertainty or for forecasts of rare and unprecedented events, statistical forecasting techniques do not suffice, and judgmental forecasting techniques are required [14,21]. Furthermore, it has been suggested that only those working closely on advanced AI technologies such as AGI may be qualified to make forecasts for such technologies [103]. However, the framework and method proposed here do not go so far as to remove all elements of statistical or data-based forecasting. Rather, we believe that all resources should be used as best as possible. Therefore, the holistic perspective focuses on judgmental techniques while using data-based statistical forecasting techniques to inform them. The framework and method are both inspired by a mixed methods approach to forecasting that uses both qualitative and quantitative judgmental methods. This mixed methods approach to the technique and the framework is another one of its strengths.

5.2. Implications for Practice

Implications for practice are straightforward and have been discussed to some degree in previous sections. However, they raise important questions about the feasibility of practical applications of this technique and framework. For one, the technique is resource intensive and requires a large number of experts for virtually all of the process. Skeptics may see the issue of expert involvement to immediately render the method inviable, and while we believe such a perspective may be extreme, the number or required experts is a credible challenge that should be addressed. There are several ways to consider soliciting experts for participation and they depend on the nature of the organization pursuing the forecast. Academic organizations may have more trouble incentivizing experts while organizations like the Partnership for AI, or the companies which comprise it, may be able to leverage member organizations' or employees' cooperation to obtain expert opinion. It is also likely that motivated and well-funded non-profit organizations (e.g., The Open Philanthropy Project) could effectively solicit expert opinion by means of paying experts appropriately for their time. Another considerable challenge is obtaining an appropriate sample of those actively working on relevant projects, and, based on the nature of the work being done, it may be desirable to intentionally collect biased samples [103]. (This would be equivalent to weighting work being conducted at certain organizations.)

Perhaps equally as costly for practical implementations of JDM and the proposed framework would be the requirement of forecasting expertise. It may be difficult to maintain full-time forecasting experts on the payroll of any organization, even one created specifically for the task of AI forecasting. Limited mappings could be developed by a single forecasting expert, and this may be sufficient for demonstrating the viability of the concept. However, the most comprehensive and accurate results would likely be realized with forecasting teams. In such teams it may be more appropriate to primarily retain forecasting experts as advisors for management and consultation while employing early career forecasters for the majority of tasks.

5.3. Implications for Research

AI forecasting is a nascent discipline, and, to date, no unifying document exists. While this work does not intend to be such a document, we do wish to draw attention to the need of the research community for one. This document presents a framework that relies on numerous techniques working harmoniously toward a single goal. Therefore, research to improve the method and framework may be most effective through analysis of some of the contributing elements. Moreover, work is also necessary which explores this framework, improvements or variations thereof, and alternate ways to incorporate judgmental forecasting methods with statistical forecasting methods and scenario analysis methods (e.g., Brundage [98]). Here we outline some suggestions of this type for future work that could be elaborated upon in the form of an AI forecasting research agenda. A structured research agenda with a coherent vision for the forecasting space could act as the sort of unifying document needed for the AI forecasting research space. (The method and framework presented here may contribute to a clear vision for the AI forecasting space, but further input is needed from others with experience in the field to iron out a unified vision.)

This study has illuminated a large number of topics that do not seem to have received appropriate attention thus far in the study of AI forecasting. Other recent work has identified some of these topics as salient [27,86], however, the previous work has not gone so far as to suggest action to motivate progress in future research. To our knowledge, no literature review exists that is equal in scope to the one presented here with respect to AI forecasting (the depth of this literature review leaves much to be desired). We believe that going forward a major priority in the study of AI forecasting is the necessity of a large number of comprehensive literature reviews for narrow topics (e.g. the many techniques discussed here) in the context of how they may be used for the tasks involved in AI forecasting. Saura et al. demonstrate an excellent example of an effective literature review for a related topic that offers a good model for such work [104]. We also see the need for a broad, comprehensive literature review—the literature review here may be a good start, but we argue that a dedicated document is desirable. These suggestions are mentioned first as they may be the lowest hanging fruit but also have significant potential for being very useful.

The literature review here found the body of existing work was lacking in studies that had compared forecasting methods. Of the studies that did, none of these compared superforecasting and none of these considered methods when used for the purpose of technological forecasting or for AI forecasting specifically. Moreover, work considering the Delphi technique found academic work assessing its viability to be lacking due to the excessive use of students rather than professionals and experts. (The Delphi technique is intended specifically for use with experts. Some studies with students have attempted forecasts of things such as college football, for which students may be considered experts, however, the vast majority of these studies did not [21].) Since the literature review was not comprehensive, a focused and more extensive effort may illuminate valuable work that has not yet been uncovered (this illustrates the possible significance that literature reviews can play). Regardless, it is clear that significant future work on methods evaluation and comparison, particularly for the viability of various forecasting techniques in the context AI forecasting, are required in order to best determine how and when the wide variety of methods are suitable in this framework and when they are suitable for AI forecasting purposes more generally. Three methods are depicted in Figure 4 as being optional inputs for JDM in the framework: the Delphi technique, prediction markets and superforecasting. It may be that one of these is indeed superior for the majority of related tasks, or, that they each can serve certain purposes in a balanced capacity to achieve the best results. Priorities for future research include comparing these three methods directly, as well as comparing the suitability of these methods in the JDM process. Such comparisons are useful both in the context of AI forecasting as well as in other contexts.

Calibration is another topic related to judgmental forecasting techniques that could be helpful for AI forecasting if better understood. Gruetzemacher et al. recently demonstrated an alternative calibration technique (i.e. naïve calibration) that demonstrates the possibility of novel calibration techniques [10]. While calibration is widely used and fundamental to superforecasting techniques, there remains little work on the topic. No work exists to confirm or assess the value that calibration training can have in forecasting, or what level of training is necessary for improving untrained experts' forecasts. Straightforward empirical studies to assess calibration training or different types of new calibration techniques for various judgmental forecasting techniques are likely very valuable. As Gruetzemacher et al. has recently shown, a substantial proportion of AI experts' forecasts are poorly calibrated, such studies to improve and better understand calibration techniques could have a quick and nontrivial return for AI forecasting efforts.

Tech mining and bibliometric mapping have a huge role to play in the proposed framework, and likely in any AI forecasting framework. Brundage mentions the use of bibliometric methods for mapping inputs [6], however, no work is known which has pursued this suggestion. While the foremost priority is likely a review of the related literature, practical work is also desperately needed. It is likely that these techniques must be used to some degree to demonstrate and/or validate the proposed method and framework, but a more extensive examination of these techniques should also be a priority. Examples of such powerful new techniques for language modeling [105], citation analysis [106] and data text mining [107] should be explored for their suitability in this topic (a literature review could likely identify even more). A large body of software also exists for the mapping of science [108], however, each flavor produces a different result. We are uncertain as to what form of results will in fact be useful for judgmental distillation, or for alternate forecasting frameworks, and this is of critical interest for an initial inquiry or for numerous simultaneous inquiries. It may be that several techniques are valid and can be shown to one expert or different experts in the JDM process. It could also be that an interactive mapping platform is most valuable for judgmental distillation in that such a platform could enable active navigation through complex maps in three dimensions. If this is the case, and if the needs for AI mapping are not met with existing software, future work could also be necessary for developing, testing and refining a tech mining based AI mapping platform. Regardless, it seems imperative that work on these topics be prioritized.

Another topic of interest is that of relevant indicators of AI progress. The literature review here discussed a growing body of work in this direction, and this work is certainly desirable moving forward. The framework proposed here takes a slightly different perspective than a substantial portion of the work toward these ends in that it suggests value in a larger number of salient indicators as opposed to a smaller number of indicators. A larger number of indicators is more realistic for the high dimensional space of AI progress, and trying to reduce progress toward broadly capable systems to a small number of vectors ignores the fundamental uncertainty of the technology that we are trying to forecast. The indicator proposed by Amodei and Hernandez met the criteria for technology indicators relatively well [69], as did some of the other metrics that have been developed for measuring AI progress [93]. Ongoing efforts toward the latter are likely sufficient at this time, but efforts to explore indicators like the former, or like the one depicted in Figure 3, while being difficult to identify, should be considered a prioritiy. There are a large range of possible outcomes and any forecasting framework must consider this.

Finally, forecasting targets are another critical topic that should be considered for future research efforts. The desiderata presented by Dafoe is an excellent ideal [7], however, in practice it can be challenging to develop targets. Structured methodologies to develop these targets are highly desirable. Work on such methods is therefore a priority given the significance of including expert forecasts for adjustment in addition to statistical forecasts. For example, workshopping techniques or interview techniques to identify and refine these targets are sorely needed. Targets that suffice may be more realistic than ideal targets, and a list of minimum criteria may be useful in addition to Dafoe's desiderata. An initial effort to expound on Dafoe's work and to examine its strengths and weaknesses could be a simple yet valuable contribution now. Also of interest are the effects of combining statistical forecasts with judgmental forecasts and aggregated expert opinion forecasts for adjustment, or the effects of combining multiple forecasts of other variations. Work exploring these effects could be examined with or without the use of domain expertise and could have significant implications on how forecasters deconstruct and delegate questions to experts in the judgmental distillation process. Furthermore, the methods of determining forecasting targets' resolution may sometimes be ambiguous and techniques are necessary for objectively resolving forecasting targets.

5.4. Challenges and Future Work

This paper only outlines and describes a new method and a new foreacsting framework. Many challenges lie ahead for continuing research on this topic. Foremost, efforts should be undertaken to evaluate and validate both the method and framework proposed here. The method may be possible to evaluate objectively in contexts other than AI forecasting, doing so may be a good step for confirming the viability of the method and the framework, however, evaluation should not be limited to toy cases. An alternate means of validation is to employ the method first in a preliminary fashion, as for demonstrating viability, and then to pursue a full-scale implementation of the method and framework. Results from the former could be used for validating and refining the technique and framework through the inclusion of calibration targets ranging from one to three years. If this was done to pilot the proposed method and framework, validation of the method would be confirmed gradually over three years. If successful early on, then further resources could be justified moving forward if performance persisted. This would also have the added benefits of improving training for AI strategy researchers and professionals and improving the scenario planning capabilities of the strategy community. If timeline forecasts were equal to or less accurate than existing methods, over short time frames, then qualitative assessment of the benefits to the planning process would have to be considered also, and the forecasting framework could be modified. Work is ongoing toward these ends. Other work should also be prioritized that assesses and validates the framework and method in other contexts which may not take as long, or, that works to refine and improve the framework and method. The framework and new method could also possibly be decomposed, evaluated and validated piece-wise to expediate the process. Regardless the path chosen, much difficult work certainly lies ahead.

6. Conclusions

The framework proposed here is not intended to be the solution for AI forecasting. Rather, it is intended to illuminate the possibility of considering a holistic perspective when addressing the unique challenges of AI forecasting. It differs from previous work in its holistic perspective and through the development of a new method for judgmental distillation of the salient features of a diverse group of forecasting techniques. By incorporating expert judgment in addition to historical and mined data, it attempts to address issues of severe uncertainty inherent in AI forecasting, while still harnessing the power of statistical methods and scenario analysis in a novel manner through the proposed framework.

There are several significant novel contributions of this work. First, the paper proposes and outlines a new method for mapping and forecasting transformative AI with judgmental distillation mapping (JDM). Second, the paper proposes and outlines a new framework for forecasting transformtive AI that builds on the new method of JDM while incorporating a variety of forecasting techniques in a holistic approach. Finally, the paper approaches the problem of forecasting in a holistic manner by incorporating many competiting methods into a single forecasting ecosystem. There are significant social implications because this method and framework, unlike any other approaches, is able to combine complex yet plausible future scenarios with a rigorous methodological foundation. In doing so, it has the potential to compel lawmakers to act on policy recommendations that in the past have seemed too unrealistic or implausible. Ultimately, the intended beneficiaries of this new approach are lawmakers constituents and all the world's citizens.

Funding: This research received no external funding.

Acknowledgments: We would like to thank Zoe Cremer, David Paradice and Christy Manning for their comments, suggestions and assistance in various aspects of the development of this text.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Roper, A.T.; Cunningham, S.W.; Porter, A.L.; Mason, T.W.; Rossini, F.A.; Banks, J. *Forecasting and Management of Technology*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
2. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. Available online: https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf/ (accessed on 21 June 2019).
3. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
4. Team. OpenAI Five. Available online: https://openai.com/blog/openai-five/ (accessed on 31 May 2019).
5. Building High-Level Features Using Large Scale Unsupervised Learning. Available online: https://icml.cc/2012/papers/73.pdf (accessed on 21 June 2019).
6. Brundage, M. Modeling progress in AI. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
7. Dafoe, A. *AI Governance: A Research Agenda*; Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
8. Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.-F.; Breazeal, C.; Crandall, J.W.; Christakis, N.A.; Couzin, I.D.; Jackson, M.O. Machine Behaviour. *Nature* **2019**, *568*, 477–486. [CrossRef]
9. Duckworth, P.; Graham, L.; Osborne, M.A. Inferring Work Task Automatability from AI Expert Evidence. In Proceedings of the 2nd Conference on Artificial Intelligence for Ethics and Society, Honolulu, HI, USA, 26–28 January 2019.
10. Forecasting Transformative AI: An Expert Survey. Available online: https://arxiv.org/abs/1901.08579 (accessed on 21 June 2019).
11. Hutter, M. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*; Springer: Berlin/Hiedelberg, Germany, 2005.
12. Minsky, M.L.; Singh, P.; Sloman, A.J. The St. Thomas common sense symposium: Designing architectures for human-level intelligence. *AI Mag.* **2004**, *25*, 113.
13. Drexler, K.E. Reframing Superintelligence. Available online: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf (accessed on 31 May 2019).
14. Armstrong, J.S. *Principles of Forecasting: A Handbook for Researchers and Practitioners*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001; Volume 30.
15. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*; OTexts: Melbourne, Australia, 2018.
16. Orrell, D.; McSharry, P. System economics: Overcoming the pitfalls of forecasting models via a multidisciplinary approach. *Int. J. Forecast.* **2009**, *25*, 734–743. [CrossRef]
17. Cooke, R. *Experts in Uncertainty: Opinion and Subjective Probability in Science*; Oxford University Press on Demand: New York, NY, USA, 1991.
18. Aspinall, W. A route to more tractable expert advice. *Nature* **2010**, *463*, 294. [CrossRef]
19. Bradfield, R.; Wright, G.; Burt, G.; Cairns, G.; Van Der Heijden, K. The origins and evolution of scenario techniques in long range business planning. *Futures* **2005**, *37*, 795–812. [CrossRef]
20. Amer, M.; Daim, T.U.; Jetter, A. A review of scenario planning. *Futures* **2013**, *46*, 23–40. [CrossRef]
21. Goodwin, P.; Wright, G. The limits of forecasting methods in anticipating rare events. *Technol. Forecast. Soc. Chang.* **2010**, *77*, 355–368. [CrossRef]
22. Rowe, G.; Wright, G. Expert opinions in forecasting: The role of the Delphi technique. In *Principles of Forecasting*; Springer: Dordrecht, The Netherlands, 2001; pp. 125–144.

23. Arrow, K.J.; Forsythe, R.; Gorham, M.; Hahn, R.; Hanson, R.; Ledyard, J.O.; Levmore, S.; Litan, R.; Milgrom, P.; Nelson, F.D.; et al. The Promise of Prediction Markets. *Science* **2008**, *320*, 877–878. [CrossRef]
24. Green, K.C.; Armstrong, J.S.; Graefe, A. Methods to Elicit Forecasts from Groups: Delphi and Prediction Markets Compared. *Foresight* **2007**, *8*, 17–20.
25. Tetlock, P.E.; Gardner, D. *Superforecasting: The Art and Science of Prediction*; Penguin Random House: New York, NY, USA, 2016.
26. Schoemaker, P.J.; Tetlock, P.E. Superforecasting: How to upgrade your company's judgment. *Harv. Bus. Rev.* **2016**, *94*, 72–78.
27. Beard, S.; Rowe, T.; Fox, J. *An Analysis and Evaluation of Methods Currently Used to Quantify Existential Risk*, under review.
28. Sanders, N.R.; Ritzman, L.P. Judgmental adjustment of statistical forecasts. In *Principles of Forecasting*; Springer: Dordrecht, The Netherlands, 2001; pp. 405–416.
29. Tversky, A.; Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **1974**, *185*, 1124–1131. [CrossRef]
30. Goodwin, P.; Wright, G. Enhancing strategy evaluation in scenario planning: A role for decision analysis. *J. Manag. Stud.* **2001**, *38*, 1–16. [CrossRef]
31. Wright, G.; Goodwin, P. Decision making and planning under low levels of predictability: Enhancing the scenario method. *Int. J. Forecast.* **2009**, *25*, 813–825. [CrossRef]
32. Lipinski, A.; Loveridge, D. Institute for the future's study of the UK, 1978–1995. *Futures* **1982**, *14*, 205–239. [CrossRef]
33. Rea, L.M.; Parker, R.A. *Designing and Conducting Survey Research: A Comprehensive Guide*; John Wiley & Sons: San Francisco, CA, USA, 2014.
34. Watts, R.J.; Porter, A.L. Innovation forecasting. *Technol. Forecast. Soc. Chang.* **1997**, *56*, 25–47. [CrossRef]
35. Porter, A.L.; Cunningham, S.W. Tech mining. *Compet. Intell. Mag.* **2005**, *8*, 30–36.
36. Phaal, R.; et al. Technology roadmapping—A planning framework for evolution and revolution. *Technol. Forecast. Soc. Chang.* **2004**, *71*, 5–26. [CrossRef]
37. Duin, P.A. *Qualitative Futures Research for Innovation*; Eburon Academic Publishers: Delft, The Netherlands, 2006.
38. Garcia, M.L.; Bray, O.H. *Fundamentals of Technology Roadmapping*; Sandia National Labs: Albuquerque, NM, USA, 1997.
39. Rip, A. Mapping of science: Possibilities and limitations. In *Handbook of Quantitative Studies of Science and Technology*; Elsevier: Amsterdam, The Netherlands, 1988; pp. 253–273.
40. Tijssen, R.J.; Van Raan, A.F. Mapping changes in science and technology: Bibliometric co-occurrence analysis of the R&D literature. *Eval. Rev.* **1994**, *18*, 98–115.
41. Nagy, B.; Farmer, J.D.; Bui, Q.M.; Trancik, J.E. Statistical basis for predicting technological progress. *PLoS ONE* **2013**, *8*, e52669. [CrossRef]
42. Mullins, C. *Retrospective Analysis of Technology Forecasting: In-Scope Extension*; The Tauri Group: Alexandria VA, USA, 2012.
43. Brynjolfsson, E.; Mitchell, T. What can machine learning do? Workforce implications. *Science* **2017**, *358*, 1530–1534. [CrossRef]
44. Van der Heijden, K.; Bradfield, R.; Burt, G.; Cairns, G.; Wright, G. *The Sixth Sense: Accelerating Organizational Learning with Scenarios*; John Wiley & Sons: San Francisco, CA, USA, 2002.
45. Perla, P.P. *The Art of Wargaming: A Guide for Professionals and Hobbyists*; Naval Institute Press: Annapolis, MD, USA, 1990.
46. Roxburgh, C. The use and abuse of scenarios. *Mckinsey Q.* **2009**, *1*, 1–10.
47. Chermack, T.J.; Lynham, S.A.; Ruona, W.E. A review of scenario planning literature. *Futures Res. Q.* **2001**, *17*, 7–32.
48. Gordon, T.J.; Helmer, O. *Report on a Long-Range Forecasting Study*; Rand Corp: Santa Monica, CA, USA, 1964.
49. Schoemaker, P.J. Scenario planning: A tool for strategic thinking. *Sloan Manag. Rev.* **1995**, *36*, 25–50.
50. Van Vliet, M.; Kok, K.; Veldkamp, T. Linking stakeholders and modellers in scenario studies: The use of Fuzzy Cognitive Maps as a communication and learning tool. *Futures* **2010**, *42*, 1–14. [CrossRef]

51. Soetanto, R.; Dainty, A.R.; Goodier, C.I.; Austin, S.A. Unravelling the complexity of collective mental models: A method for developing and analysing scenarios in multi-organisational contexts. *Futures* **2011**, *43*, 890–907. [CrossRef]
52. List, D. Scenario network mapping. *J. Futures Stud.* **2007**, *11*, 77–96.
53. Inayatullah, S. Causal layered analysis: Poststructuralism as method. *Futures* **1998**, *30*, 815–829. [CrossRef]
54. Axelrod, R. *Structure of Decision: The Cognitive Maps of Political Elites*; Princeton University Press: Princeton, NJ, USA, 2015.
55. Tolman, E.C. Cognitive maps in rats and men. *Psychol. Rev.* **1948**, *55*, 189–208. [CrossRef]
56. Jetter, A.J.; Kok, K. Fuzzy Cognitive Maps for futures studies—A methodological assessment of concepts and methods. *Futures* **2014**, *61*, 45–57. [CrossRef]
57. Papageorgiou, E.I. *Fuzzy Cognitive Maps for Applied Sciences and Engineering: From Fundamentals to Extensions and Learning Algorithms*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 54.
58. List, D. *Scenario Network Mapping: The Development of a Methodology for Social Inquiry*; University of South Australia: Adelaide, Australia, 2005.
59. Michie, D. Machines and the theory of intelligence. *Nature* **1973**, *241*, 507–512. [CrossRef]
60. Grace, K. AI Timeline Surveys. Available online: https://aiimpacts.org/ai-timeline-surveys/ (accessed on 31 May 2019).
61. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **2018**, *62*, 729–754. [CrossRef]
62. Zhang, B.; Dafoe, A. *Artificial Intelligence: American Attitudes and Trends*; University of Oxford: Oxford, UK, 2019.
63. Baum, S.D.; Goertzel, B.; Goertzel, T.G. How long until human-level AI? Results from an expert assessment. *Technol. Forecast. Soc. Chang.* **2011**, *78*, 185–195. [CrossRef]
64. Müller, V.C.; Bostrom, N. Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental Issues of Artificial Intelligence*; Springer International Publishing: Basel, Switzerland, 2016; pp. 555–572.
65. Armstrong, S.; Sotala, K. How we're predicting AI–or failing to. In *Beyond Artificial Intelligence*; Springer: Pilsen, Czech Republic, 2015; pp. 11–29.
66. Armstrong, S.; Sotala, K.; hÉigeartaigh, S.Ó. The errors, insights and lessons of famous AI predictions–and what they mean for the future. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 317–342. [CrossRef]
67. What Do We Know About AI Timelines? Available online: https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/ai-timelines (accessed on 31 May 2019).
68. What Should We Learn from Past AI Forecasts? Available online: https://www.openphilanthropy.org/focus/global-catastrophic-risks/potential-risks-advanced-artificial-intelligence/what-should-we-learn-past-ai-forecasts (accessed on 31 May 2019).
69. AI and Compute. Available online: https://openai.com/blog/ai-and-compute/ (accessed on 31 May 2019).
70. Interpreting AI Compute Trends. Available online: https://aiimpacts.org/interpreting-ai-compute-trends/ (accessed on 31 May 2019).
71. Reinterpreting "AI and Compute". Available online: https://aiimpacts.org/reinterpreting-ai-and-compute/ (accessed on 31 May 2019).
72. Measuring the Progress of AI Research. Available online: https://www.eff.org/ai/metrics (accessed on 31 May 2019).
73. Trends in Algorithmic Progress. Available online: https://aiimpacts.org/trends-in-algorithmic-progress/ (accessed on 31 May 2019).
74. Constantin, S. Performance Trends in AI. Available online: https://srconstantin.wordpress.com/2017/01/28/performance-trends-in-ai/ (accessed on 31 May 2019).
75. AI Metrics Data. Available online: https://raw.githubusercontent.com/AI-metrics/master_text/master/archive/AI-metrics-data.txt (accessed on 31 May 2019).
76. Adams, S.; Arel, I.; Bach, J.; Coop, R.; Furlan, R.; Goertzel, B.; Hall, J.S.; Samsonovich, A.; Scheutz, M.; Schlesinger, M. Mapping the landscape of human-level artificial general intelligence. *AI Mag.* **2012**, *33*, 25–42. [CrossRef]
77. Goertzel, B. *The AGI Revolution: An Inside View of the Rise of Artificial General Intelligence*; Humanity+ Press: Los Angeles, CA, USA, 2016.
78. Goertzel, B. *Ten Years to the Singularity If We Really Really Try*; Humanity+ Press: Los Angeles, CA, USA, 2014.

79. Gruetzmacher, R.; Paradice, D. Alternative Techniques for Mapping Paths to HLAI. *arXiv* **2019**, arXiv:1905.00614.
80. Computing Community Consortium (CCC) (Ed.) Townhall: A 20-Year Roadmap for AI Research. In Proceedings of the 33nd Annural Conference for the Association of the Advancement of Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
81. Bostrom, N. *Superintelligence*; Oxford University Press: Oxford, UK, 2014.
82. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741 (accessed on 31 May 2019).
83. Barrett, A.M.; Baum, S.D. Risk analysis and risk management for the artificial superintelligence research and development process. In *The Technological Singularity*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 127–140.
84. Barrett, A.M.; Baum, S.D. A model of pathways to artificial superintelligence catastrophe for risk and decision analysis. *J. Exp. Theor. Artif. Intell.* **2017**, *29*, 397–414. [CrossRef]
85. Baum, S.; Barrett, A.; Yampolskiy, R.V. Modeling and interpreting expert disagreement about artificial superintelligence. *Informatica* **2017**, *41*, 419–428.
86. Avin, S. Exploring Artificial Intelligence Futures. *J. AI Humanit.*. Forthcoming.
87. Parson, E.A. *What Can You Learn from A Game? Wise Choices: Games, Decisions, and Negotiations*; Harvard Business School Press: Boston, MA, USA, 1996.
88. Davis, P.K. *Illustrating a Model-Game-Model Paradigm for Using Human Wargames in Analysis*; RAND National Defense Research Institute: Santa Monica, CA, USA, 2017.
89. Fernández-Macías, E.; Gómez, E.; Hernández-Orallo, J.; Loe, B.S.; Martens, B.; Martínez-Plumed, F.; Tolan, S. A multidisciplinary task-based perspective for evaluating the impact of AI autonomy and generality on the future of work. *arXiv* **2018**, arXiv:1807.02416.
90. Evaluation of General-Purpose Artificial Intelligence: Why, What & How. Available online: http://dmip.webs.upv.es/EGPAI2016/papers/EGPAI_2016_paper_9.pdf (accessed on 31 May 2019).
91. Hernández-Orallo, J. AI Evaluation: Past, Present and Future. *arXiv* **2014**, arXiv:1408.6908.
92. Martínez-Plumed, F.; Avin, S.; Brundage, M.; Dafoe, A.; hÉigeartaigh, S.Ó.; Hernández-Orallo, J. Accounting for the neglected dimensions of ai progress. *arXiv* **2018**, arXiv:1806.00610.
93. Martínez-Plumed, F.; Hernández-Orallo, J. Analysing Results from AI Benchmarks: Key Indicators and How to Obtain Them. *arXiv* **2018**, arXiv:1811.08186.
94. Hernández-Orallo, J. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2017.
95. Riedl, M.O. The Lovelace 2.0 test of artificial intelligence and creativity. In Proceedings of the 29th AAAI Conference on Artificial Intelligence Workshops, Austin, TX, USA, 25–26 January 2015.
96. Hernández-Orallo, J.; Baroni, M.; Bieger, J.; Chmait, N.; Dowe, D.L.; Hofmann, K.; Martínez-Plumed, F.; Strannegård, C.; Thórisson, K.R. A new AI evaluation cosmos: Ready to play the game? *AI Mag.* **2017**, *38*, 66–69. [CrossRef]
97. Castelvecchi, D. Tech giants open virtual worlds to bevy of AI programs. *Nat. News* **2016**, *540*, 323. [CrossRef]
98. Brundage, M. *Responsible Governance for Artificial Intelligence: An Assessment, Theoretical Framework, and Exploration*, 2018, Unpublished.
99. Gruetzmacher, R.; Paradice, D. Mapping the Paths to AGI. In Proceedings of the 12th Annual Conference on Artificial General Intelligence, Shenzhen, China, 6–9 August 2019.
100. Clune, J. AI-GAs: AI-Generating Algorithms, an Alternate Paradigm for Producing General Artificial Intelligence. *arXiv* **2019**, arXiv:1905.10985.
101. List, D. *Scenario Mapping: A User's Manual*; Original Books: Adelaide, Australia, 2006.
102. Gruetzemacher, R. Rethinking AI Strategy and Policy as Entangled Super Wicked Problems. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; p. 122.
103. There's No Fire Alarm for Artificial General Intelligence. Available online: https://intelligence.org/2017/10/13/fire-alarm/ (accessed on 31 May 2019).
104. Saura, J.R.; Palos-Sánchez, P.; Cerdá Suárez, L.M. Understanding the digital marketing environment with KPIs and web analytics. *Future Internet* **2017**, *9*, 76. [CrossRef]

105. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

106. Cohan, A.; Ammar, W.; van Zuylen, M.; Cady, F. Structural Scaffolds for Citation Intent Classification in Scientific Publications. *arXiv* **2019**, arXiv:1904.01608.

107. Saura, J.R.; Bennett, D.R. A Three-Stage method for Data Text Mining: Using UGC in Business Intelligence Analysis. *Symmetry* **2019**, *11*, 519. [CrossRef]

108. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]

big data and
cognitive computing

Article

Peacekeeping Conditions for an Artificial Intelligence Society

Hiroshi Yamakawa [1,2,3]

[1] The Whole Brain Architecture Initiative, a Specified Non-Profit Organization, Nishikoiwa 2-19-21, Edogawa-ku, Tokyo 133-0057, Japan; ymkw@wba-initiative.org
[2] The RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-Chome Mitsui Building, 15th Floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan
[3] Dwango Co., Ltd., KABUKIZA TOWER, 4-12-15 Ginza, Chuo-ku, Tokyo 104-0061, Japan

Received: 18 May 2019; Accepted: 20 June 2019; Published: 22 June 2019

Abstract: In a human society with emergent technology, the destructive actions of some pose a danger to the survival of all of humankind, increasing the need to maintain peace by overcoming universal conflicts. However, human society has not yet achieved complete global peacekeeping. Fortunately, a new possibility for peacekeeping among human societies using the appropriate interventions of an advanced system will be available in the near future. To achieve this goal, an artificial intelligence (AI) system must operate continuously and stably (condition 1) and have an intervention method for maintaining peace among human societies based on a common value (condition 2). However, as a premise, it is necessary to have a minimum common value upon which all of human society can agree (condition 3). In this study, an AI system to achieve condition 1 was investigated. This system was designed as a group of distributed intelligent agents (IAs) to ensure robust and rapid operation. Even if common goals are shared among all IAs, each autonomous IA acts on each local value to adapt quickly to each environment that it faces. Thus, conflicts between IAs are inevitable, and this situation sometimes interferes with the achievement of commonly shared goals. Even so, they can maintain peace within their own societies if all the dispersed IAs think that all other IAs aim for socially acceptable goals. However, communication channel problems, comprehension problems, and computational complexity problems are barriers to realization. This problem can be overcome by introducing an appropriate goal-management system in the case of computer-based IAs. Then, an IA society could achieve its goals peacefully, efficiently, and consistently. Therefore, condition 1 will be achievable. In contrast, humans are restricted by their biological nature and tend to interact with others similar to themselves, so the eradication of conflicts is more difficult.

Keywords: autonomous distributed system; conflict; existential risk; distributed goals management; terraforming; technological singularity

1. Introduction

Emergent technology is continually advancing because of its many benefits for humankind. However, technology is not always used for good. As a result, the number of people who have destructive offensive capabilities are increasing. These trends enhance existential risks such as deliberate misuse of nanotechnology, nuclear holocaust, and badly programmed superintelligence [1]. Specifically, the existence of a small number of persons whose aim is to use AI for destructive purposes has the potential to have an enormous impact on humanity. Suspicion between nations has the potential to cause a disastrous war [2]. This irreversible change is also called the "threat of universal unilateralism" [3], and this sufficiently high existential risk could explain Fermi's paradox: "humanity has no experience of contact with civilized extraterrestrials, compared to their potentially high likelihood of existence" [4].

In many cases, such an abuse of advanced technology is motivated by conflicts in societies, but eliminating all conflict is impossible. Today, the accelerating innovation by, the artificial intelligence (AI) and the recruiting human resource for that are the kinds of major factors of competition when nations and organizations seek to gain supremacy [5,6]. From this background, a human society equipped with advanced technology cannot sustain itself without keeping the peace despite various conflicts. Humankind has made many efforts to maintain peace, including the creation of institutions and organizations such as the United Nations and International Law and Peace Keeping Operation, and their effects have been observed. However, they have been unable to eradicate disputes, wars, and conflicts. Thus, maintaining peace among human societies using only human efforts remains a challenge.

AI will gradually surpass human intelligence, and human-level artificial general intelligence is estimated to be created by 2100 [7]. In general, this unpredictable change is feared due to various dangers [8–12], but this change will provide us with an opportunity to eradicate disputes, conflicts, terrorism, and wars. Peace in human society can be achieved through appropriate interventions by advanced artificial intelligence, rather than by human effort. Figure 1 shows an example of an ecosystem in which an AI system built as a society of intelligent agents (IAs) supports human society. In this example, basically, the AI society observes the values of individuals and/or groups and provides them with benefits. Simultaneously, based on the common values of all of humanity, IAs arbitrate conflicts and contradictions that exist in human society. AIs also act to persuade and educate individuals and groups.

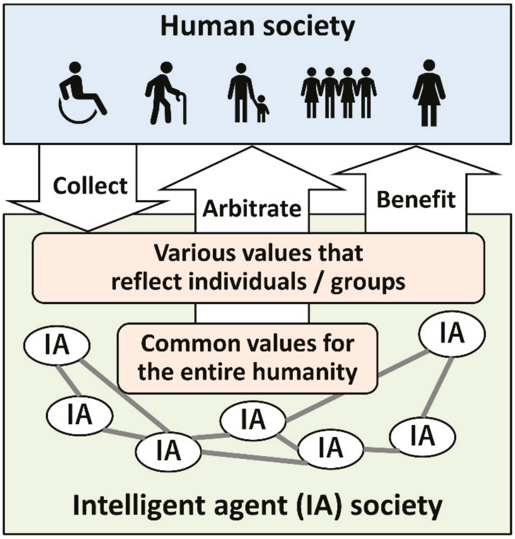

Human society
Collect
Arbitrate
Benefit
Various values that reflect individuals / groups
Common values for the entire humanity
IA
IA
IA
IA
IA
IA
IA
Intelligent agent (IA) society

Figure 1. Example ecosystem consisting of human society and artificial intelligence (AI) society. Note: IA denotes intelligent agent which contains AI.

For the AI system to keep the peace in human society, at least the following three conditions must be satisfied:

(1) Condition 1: the AI system is operating continuously and is stable.

(2) Condition 2: the AI system has an intervention method that maintains peace in human societies based on a common value or values.

Additionally, as a premise, the following conditions are required. These conditions involve the sustainable development goals (SDGs), which are a collection of the wisdom of many people, and are potential candidates for common values of humankind in the near future.

(3) Condition 3: a minimum common value must exist that can be agreed upon across human society.

In this study, a thought experiment was conducted to investigate the first condition. The research question is "Is it possible to build an AI system that operates continuously and stably?". The reason this argument is necessary is because an AI system that inevitably creates a society of autonomous decentralized agents can be destabilized by the occurrence of competition, as well as by human society. For an AI system that is becoming increasingly fast, the range of operations that can be controlled by humans decreases. Therefore, the AI system must be able to operate stably without human assistance.

In the next chapter, I explain the setting for a thought experiment in which an AI system is responsible for the execution of terraforming. This mission can be currently readily agreed upon as a common goal by all human beings. AI systems need to be able to react quickly to various situations and must be robust against threats of destruction and failure. For these reasons, the AI system should be a team of distributed autonomous intelligent agents (IAs). In Section 3, the occurrence of contradictions, competition, and conflict in a society of autonomous decentralized IAs is investigated. Even in this case, peace can be maintained if all IAs think that the other IAs share similar goals; however, several obstacles exist to realizing this ideal situation. In Section 4, I argue that a distributed goal-management system for the IA society can be constructed to support sharing goals among agents. By introducing this system, conflicts in IA society can be arbitrated and peace can be realized. In Section 5, I discuss the reasons why it is difficult for human society to maintain peace by itself in comparison with an AI society. In Section 6, the first argument is that IAs can be comrades for human beings, unlike other animals. Further, I argue that unlike human society, majority decision-making does not make sense for an IA society. The major conclusions are finally summarized in Section 7.

2. Thought Experiment Settings

Before explaining the terraforming that is the subject of the thought experiment in the following sections, a trivial example in which resource competition creates peaceful cooperation rather than conflict is explained. For example, in the case where a deep reinforcement learning agent [13] runs searches in parallel for a parameter that achieves the highest score in a certain game task, the agent plans how and in what order to conduct a number of experiments. During the process, no parameter fights another over finite computational resources, and thus there can be no problematic situation that requires resolution. Due to the clarity of common goals, and due to an absence of local goals in each parameter set, there is no competition among parameter sets (although such virtual competition paradigms can be possible). This means that there can be no struggle under shared common goals and unified management.

This section is divided into subheadings. It provides a concise and precise description of the experimental results, their interpretation, and the experimental conclusions that can be drawn.

2.1. Intelligent Agent Society for Terraforming

In the above example in which no agent has local goals, no competition can occur. As the setting of the thought experiment in this article, it is assumed that a number of IAs that can proliferate themselves have been sent to an unknown planet. The background of this setting is that the IAs are dispatched from Earth by humans, and charged with the mission of remaking the environment of the planet in preparation for human migration. Thus, the goal of these IAs is transforming the planet into being human-habitable; that is, terraforming. The Invincible (first published in 1945) by Stanisław Lem [14] and Code of the Lifemaker by James P. Hogan [15] are famous fictional examples of this kind of scenario.

This science fiction-like setting is introduced for two reasons. The first is to simplify the structure of the struggle in human society. The second is to reduce the influence of biased thinking caused by personal history. The procedure of providing an objective function to an agent is standard in artificial intelligence research, and can be easily discussed. In addition, I think that expanding human habitats to other planets is an effective method to increase the survivability of humanity.

In the following sections, the scenario that the IA society avoids disputes and war and maintains peace, despite technological progress causing conflicts among that society, is described.

2.2. Autonomous Distributed Agents: For the Survival of the Group

Fortunately, a group of IAs have landed on a planet. They have secured resources including energy and have successfully survived for the time being. To achieve their common goal, which is the terraforming of this new planet, they begin striving and working toward it.

For the following reasons, each member of the group must be spatially distributed and autonomous [16] because, firstly, each IA should adapt to the surrounding environment to respond quickly to what is in front of its eyes with limited information processing capabilities. The second reason is concerned with the robust survivability of the group. The hardware of individual IAs is constantly exposed to various environmental factors and may be destroyed. To secure the survival of the group, therefore, IAs have to be highly autonomous and spatially dispersed.

2.3. Physical Composition of the IAs and Their Group

It is assumed that the hardware of each IA is a set of physical devices for memory, communication, computation, sensory inputs, locomotion, manipulation, and so forth. It is further assumed that hardware for new members is produced/reproduced in manufacturing plants, in a system that is similar to that of social insects. Unlike living organisms, though, reproduced IAs do not need to be similar to the producing IAs, as they are manufactured solely based on their design specifications. Stored data such as memory, programs, and knowledge, which arguably compose the essential substance of the IAs, are realized as software, and the dependency on their hardware can be relatively low.

Thus, the essential substance of IAs, which is their software, does not need a specific physical body and is able to wander among many bodies. Even the preservation and maintenance of the software of a specific IA are low priority because electronic data can be stored easily and restored at will. Additionally, when an IA reboots another IA, they do not need to be similar.

3. Development and Conflict in an IA Society

IAs work as an organization by communicating information, such as goals, to each other. Formation of a group leads to cooperation and division of labor within it, which contributes to efficiency in achieving their common purpose. By sharing knowledge about their environment and developing knowledge including science and technology, their efficiency continually. Closer relationships between the IAs enable useful collaborations toward their goals, but also increase conflicts.

3.1. Diversification and Fixation of Local Values: Emergence of Survival Instinct

All IAs contain a distributed autonomous system with common goals, and the members are required to retain the goals and maintain activities toward achieving them. This means all the IAs must hold the common goals individually and in a distributed manner. Each IA derives sub-goals from the common goals or from an assigned part of the common goals (target-means decomposition) in response to its environment, and builds local values as a network of sub-goals. Each IA forms specialized local values, depending on its body, tasks, and the local environment (Figure 2). This area of research is referred to as cooperated multi-agent planning (MAP), and a large amount of accumulated work on the subject has been published [17]. As each IA changes behavior by learning, and as the number of IAs increases, the coordination between them becomes more difficult.

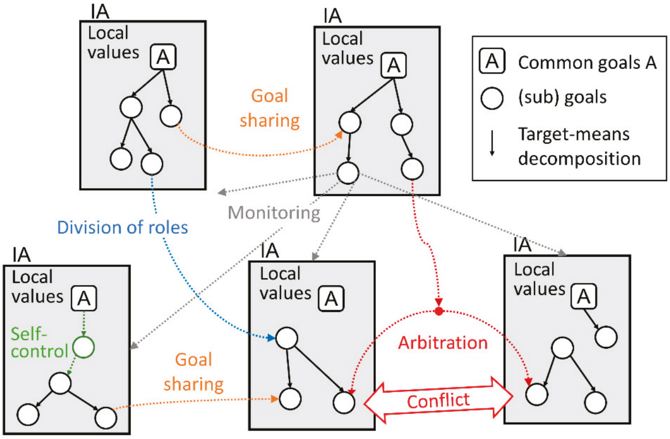


Figure 2. Sharing goals in an IA group: various IAs generate different sub-goals and specialize themselves to those sub-goals in response to their environment, bodies, and tasks. Each IA conducts target-means decomposition, arbitration, and, at times, checks consistency. They also carry out, in cooperation with other IAs, task allotment, goals sharing, arbitration, monitoring, and so forth.

Individual IAs carry out activities toward concrete sub-goals within a certain time frame. Too frequent changing of the sub-goals makes it difficult for them to solve the current problems. Thus, other sub-goals emerge—stabilization of local values. With a slightly longer time frame, self-preservation of their hardware also emerges as a sub-goal because frequent breakdowns, or shutdowns, of the hardware impairs their usefulness.

Maintenance of the IAs' software does not emerge as an overly important sub-goal because, basically, the programs of these IAs can be stored, rebooted, copied, and transferred at low cost. There is the risk that they may be destroyed by accidents, attacked, or manipulated by enemies before rebooting. Additionally, the risk exists that a proper information environment might not be available when rebooting is needed. However, if some specific program is useful from the viewpoint of the common goals, the IA society tries to secure preservation and rebooting of that program.

In case the rebooted IA needs to catch up to a change in the social situation, the AI society simply provides it with a learning period. If a rebooted IA is forgotten by the others and cannot serve IA society, that one does not need to be rebooted, and if wrongly rebooted, it is immediately shut down. Almost all the goals of IAs have terminating conditions (ending with the fulfillment of the purpose or with a judgment of infeasibility) [18]. However, the survival instinct, including self-preservation, resource acquisition, and knowledge acquisition, is always a sub-goal as long as each IA exists. Regardless of initial goals, any advanced intelligence generates sub-goals related to a survival instinct, and it sometimes becomes excessively predominant; this is called "instrumental convergence" [10].

3.2. Confident Sharing of Common Goals Is Difficult

When all IAs share mutually believable local values derived from common goals, no inconsistency or struggle will exist between IAs, and all IAs in the society can pursue common goals peacefully, efficiently, and consistently. The ideal situation is, in other words, that every agent can believe that "all other agents intend socially acceptable goals".

However, as mentioned above, when various goals are generated diversely and dynamically in each IA, different local values will be developed among them. Therefore, it is required for the cooperation and division of labor between different IAs to not only share goals, but also to be mutually confident about the shared goals. This would correspond to the establishment of a trust relationship in the contract.

It is assumed that each IA is programed to act in good faith; this means IAs do not pretend, lie, or betray.

Because the AI system is designed as distributed autonomous IAs, an IA needs to be able to do the following to act ideally for the social good:

(1) commit to socially accepted goals,
(2) send and receive goals as information to and from other IAs, and
(3) understand goals received from other IAs.

Here, "socially acceptable goals" means that the goals contribute to common goals and do not conflict with any other IA's local values in practice.

I think that a society constructed by individuals with different local values has a potential risk of conflict. Therefore, some common goals must be shared that are on levels beyond those local values to establish a single, orderly society.

There are three obstacles to achieving the above ideal situation:

(1) Communication channel problem

It is assumed that communication paths between IAs for sharing common goals as information are built in and shared with all IAs in the design stage. However, communication channels among IAs are not always stable and may be disrupted at times. According to Brewer's CAP Theorem (This theorem states that it is impossible for a distributed data store to simultaneously provide more than two out of the consistency, availability and partition tolerance) [19], when securing availability and partition tolerance in a distributed system, a delay in sharing information must be accepted.

(2) Comprehension ability problem

This problem is caused by the limitations of each IA's comprehension ability. Here, this ability means the capacity of an IA to derive sub-goals from received goals and to act on received goals. Even if the shared goals are formally identical, differences in IAs lead to different comprehension. For example, different IAs have different designs and appearances (body, experience/knowledge, capacity, etc.).

(3) Computational complexity problem

Suppose one IA overcomes the above two problems and understands the other IAs' goals. Even in this case, the following processing is required to avoid substantial conflicts. First, in the IA's own environment, all the goals held as their own local values will check for conflicts with other IA goals. Next, if a contradiction is detected, the IA needs to change its own local value so that the contradiction does not occur in view of the higher priority goals. In some cases, an IA may need to determine that it needs to request another IA to adjust its goals. This type of processing requires a considerable computational cost.

3.3. Birds of a Feather Flock Together: Agent Society for Terraforming

Given the problems mentioned above, it is difficult for IAs to share goals in a workable manner. However, if the pre-designed appearance is similar between IAs, they can infer that they have similar goals because the IAs' goals and appearances are probably governed by the same design information. If the circumstances are similar between IAs, their interests also tend to be similar. In short, when similar people gather together, the possibility of sharing goals is increased (Figure 3).

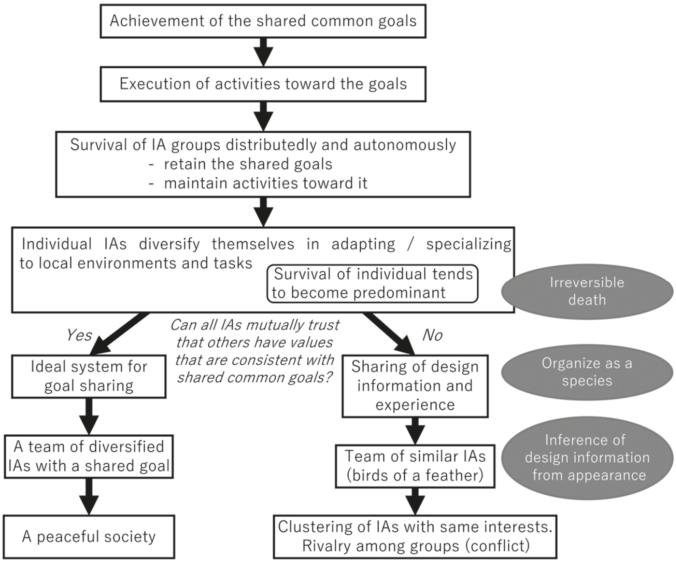

Achievement of the shared common goals
Execution of activities toward the goals
Survival of IA groups distributedly and autonomously
- retain the shared goals
- maintain activities toward it
Individual IAs diversify themselves in adapting / specializing to local environments and tasks
Survival of individual tends to become predominant
Irreversible death
Can all IAs mutually trust that others have values that are consistent with shared common goals?
Yes
No
Ideal system for goal sharing
A team of diversified IAs with a shared goal
A peaceful society
Sharing of design information and experience
Organize as a species
Team of similar IAs (birds of a feather)
Inference of design information from appearance
Clustering of IAs with same interests. Rivalry among groups (conflict)

Figure 3. The branching point of peace and conflict.

A highly homogenous team of IAs will have few conflicts. They can cooperate and divide labor efficiently, and that makes them advantageous compared with other teams. Like a flock of birds, survival probability is increased for individual members. This leads to achieving another sub-goal that they should pursue as a team: survival of the team. For an individual in a flock, the more its local values reduce conflict with other members, the better the chance it has of surviving, which promotes standardization across the entire flock.

For these reasons, IAs will tend to form highly homogeneous teams. In other words, “Birds of a feather flock together”. However, this often exposes a weakness: homogeneity makes the flock susceptible to environmental changes.

3.4. Conflict Between Groups

Within each group, similar local values are shared by the members, but they differ from those of other groups. Because resources in the world are limited, the effort to acquire them causes conflicts of interest between individuals and between groups. Similarly, divisions in IA groups will tend to cause a state of conflict between groups.

When confrontation deepens, an IA in one group perhaps ignores, disfavors, or blocks opponent IAs. Contrarily, the same IA provides preferential treatment and increased communication to members of the same group.

As already described in the introduction, if the worst should happen, a struggle could imperil the entire society. However, even in the preliminary stages of such a conflict, each IA will expand activities of attacking and defending against opposing IA groups, causing the problem of diminishing allocation of resources to the original common goals.

4. Peace of IA Society Maintained by a Distributed Goal Management System (DGMS)

A distributed goal management system (DGMS) should be introduced to make the IA society peaceful. The technological foundation of DGMS has progressed in the field of multi-agent planning (MAP) since the beginning of 90’s, as reported in a previous survey article [17]. By using DGMS, each IA coordinates its local values with other IAs’ values through dialogue, and often an individual IA needs to execute the tasks it faces in real time. Various technical issues must be overcome to realize the specifications required for practical DGMS, and it is necessary to promote research on MAP and related fields to make DGMS a reality.

If the goals of all IAs are coordinated to be socially acceptable, there is no conflict in IA society. However, communication channel problems, comprehension ability problems, and computational complexity problems are preventing this from being realized. To overcome these problems, DGMS should have the additional three functions listed below.

4.1. Normal Responses

When two IAs conflict with each other due to different sub-goals and actions derived therefrom, another appropriate IA arbitrates (or mediates) between them. A third IA estimates the importance and validity of the sub-goals of both by considering consistency and contribution to the common goals. Both IAs must comply with the ruling (Figure 2).

In cases in which the local value of an IA lacks consistency with the common goals and consistency cannot be restored by the calculation of the IA itself, another appropriate IA recalculates the sub-goals and assigns them.

4.2. Emergency Responses

Emergency responses are necessary because normal responses need time for communication and calculation. These will consist of suspending actions that are based on questionable sub-goals and even shutting down the IA temporarily for safety.

The method for detecting danger in each IA as the premise for taking these kinds of measures is as follows. First, one IA monitors the local values of many IAs and finds any sub-goals that are inconsistent with the common goals and might be a source of conflict. Second, each IA checks the consistency between its local values and common goals. This is the self-restraint of IAs (Figure 3).

This crisis management is a kind of traditional safety design technology (e.g., safe operation of aircraft).

4.3. Task Assignment in Consideration of Comprehension Ability

Due to the comprehension ability problem, the level of understanding of IAs varies depending on differences in their design and/or appearance. However, understanding goals is necessary for many goal-related processes such as execution, target-means decomposition, arbitration, and monitoring. Thus, appropriate assignment of roles to each IA by considering its comprehension ability will be an important technical consideration in designing DGMS.

If ideal and practical DGMS can be built by overcoming obstacles, the situation in which every IA thinks that "all other IAs intend socially acceptable goals" can be maintained. In that situation, IA society can achieve goals peacefully, efficiently, and consistently.

5. What Prevents Peace in Human Society?

Despite our capability to share common goals through language, humans cannot stop fighting with each other. Much of the reason for this might lie in the biological constraints to which humans are subjected. By comparing IA society to human society, the reasons that human society divides into many rival groups resulting in conflicts were considered.

5.1. Irreversible Death of Living Organisms

Death is an irreversible and inescapable event for all organisms. Humans cannot delay or switch on and off our biological activities at will. For this reason, each individual organism has to cling to life. In the future, the realization of hibernation technology may reduce this fear, but our human brain cannot escape from the fear that something could go wrong, and we might not wake up. People in hibernation might also fear that they will be forgotten by society (e.g., Rip Van Winkle [20]).

Because available resources in the world are limited, increasing numbers of individuals obsessed with survival will inevitably cause competition for resource acquisition and become an origin of conflict.

An important part of IAs is their software, which can be stored, rebooted, copied, and transferred at a low cost as described in Section 3.1. Therefore, the degree to which IAs cling to their lives may be much lower than that of living organisms.

5.2. Struggle Between Evolutionarily United Species

Evolution is a search algorithm, and it can expand the possibilities of organisms through copying, mating, and mutating individuals. The phenotypes that can survive in the environment are extremely narrow in the space of innumerable genome combinations [21]. Therefore, to ensure the survival probability of offspring born by mating, it is necessary to form a species that is a homogeneous group.

Surviving as a species requires a population above a certain number, which is called the minimum viable population (MVP) [22]; otherwise, the diversity of the genes within the species decreases and it becomes vulnerable to extinction. For this reason, individuals of a species share the same fate, and they sometimes help other members of the same species [23]. However, this leads to competition over resources among species that sometimes develops into conflict, as seen in invasive species [24].

IAs are constructed based on their specifications, and they can produce completely different offspring. Therefore, there is no incentive to increase similar mates for reproduction. The situation in which species compete for resources becomes unrealistic. In the future, gene editing technology could realize the free design of living organisms [25]. After that, humans may not need to fight to maintain the species.

5.3. Estimate Goal Similarity Based on Appearance

In the case of the IA, there is no need to be suspicious that it uses the same communication devices and shares common goals, except with regard to their failure or hacking. Living organisms sometimes validate agreement of design information using chemical interaction. However, for humans, other agents' goals are inferred from their appearance or from shared experiences. Therefore, human beings tend to be sympathetic to organisms or objects that have a similar appearance. Due to this nature of human beings, it is thought that intelligent robots like human beings will greatly affect human emotions, and there are concerns regarding various problems arising from this [26].

Human beings can communicate their goals to others through language, but they cannot know whether the other intends to commit to these goals. To secure that point, modern society uses a legally effective contract. In this case, it is premised that the people on both sides have the ability to understand the contents of the contract. However, from the viewpoint of one organization, it is not possible to confirm the intention of another organization. Stemming from this lack of confidence, almost every nation has military capabilities that are based on "offensive realism" [2].

5.4. Section Summary

As humans are also living beings, they have an individual survival instinct for avoiding irreversible death, and they try to care for members of the same species to keep the species alive. In particular, humans tend to infer the goals or intentions of other persons from their visual appearance. If the same language is used, cooperation in the group is much easier. For this reason, even in humans, the tendency of "birds of a feather flock together" is particularly strong. Therefore, competition between groups with different local values inevitably occurs.

6. Discussion

6.1. As a Comrade

The mechanisms of collaboration of non-human animals are mostly determined genetically [23]. However, from the time of evolution to homo sapiens, the scope of our recognition of a comrade began to expand, and is now expanding to all humanity through language and education [27]. The scope recognized as comrades will be extended to include intelligent machines in the future.

According to the Salient Value Similarity (SVS) model [28], whether a person is trustworthy or not depends on whether the person seems to share the same stable and consistent goals, and has the capability and enthusiasm to pursue them, from the perspective of the observer. Again, the main condition for peaceful coexistence is whether every agent can believe that "all other agents intend socially acceptable goals". In this sense, IAs with highly advanced AI will be able to share goals with humans, and have the capacity and enthusiasm to pursue them as well. Considering this, IAs will have the opportunity to become trustable comrades, more so than other intelligent animals on earth. If such an advanced AI can share a relatively wide range of values with them through the education given by surrounding people, like foster children, then it is possible that AI may become a trusted comrade.

6.2. Significance of Majority Decision

A decision by majority rule is a prevalent social decision method in human society. However, in the case of IAs whose programs are indefinitely duplicable, it is pointless to count the number of software units that agree with an opinion. Conversely, imbuing hardware with the right to vote can be somewhat meaningful, but perhaps the hardware has no opinion.

In contrast, the individuality of each living organism is of supreme importance because each software and hardware is tightly coupled. Thus, in social decisions made in human groups, , appropriate to distribute voting rights with the same weight to each person from a utilitarian perspective [29].

As for social decisions in IA society, the main point of interest is how to achieve common goals. Therefore, it is desirable that the IA group collect diverse experiences and abilities to produce diverse planning alternatives [30]. It is also desirable for the IA groups to be able to predict the degree of contribution of each alternative to the common goals as accurately as possible and, eventually, decide the best action for attaining these goals.

7. Conclusions

The development of emergent technology will increase the risk to human existence. Therefore, maintaining peace in human society by overcoming various conflicts is becoming an urgent issue. Historically, human society has not achieved full peace through its own efforts alone. Therefore, it is worthwhile to explore the possibility of realizing peace in human society through the intervention of advanced AI systems. Three conditions are assumed to be needed to realize this situation: There are minimum common values that can be agreed across society (condition 3), advanced AI systems can intervene to keep the peace of human society based on these common values (condition 2), and an AI system exists that can work stably and continuously (condition 1).

In this paper, an AI system that satisfies condition 1 was investigated. A part of the system may potentially be destroyed, so a robust IA society should be a team of autonomous and distributed IAs. A common value of humanity is shared among all IAs. Individual IAs would decompose common goals and derive means so that they can contribute to the advancement of common values. Each IA would diversify its activities to effectively divide tasks among them all. In order to adapt to their local environment, IAs would usually hold, as their local values, sub-goals derived from common goals.

There are a wide variety of local values and competition for available resources will create a competitive situation for IA comrades. It is an advantage that competition leads to an increase in capacity to achieve a common goal. However, if the effort towards merely winning the competition increases, cooperation is lost, and devastating struggles occur, creating obstacles to achieving common goals.

The ideal situation is one in which every agent believes that "all other agents intend socially acceptable goals". Here, "socially acceptable goals" means that the goals contribute to common goals and do not conflict with any other IA's local values in practice. Under such circumstances, the IA society can achieve goals peacefully, efficiently, and consistently. Communication channel problems, comprehension ability problems, and computational complexity problems exist, however, that may impede the realization of ideal situations. In an IA society based on a computer, it seems possible to design a DGMS that maintains the local values of distributed IAs.

Conversely, humans are biologically constrained. Irreversible death strengthens our survival instincts, and human beings need to maintain our species through reproduction. Humans must also distinguish their mates by appearance. For these reasons, similar people gather and become more likely to form a party. If people divide into groups with similar values and compete for resources, this can be a major cause of conflict.

I assumed that building a universal AI system to arbitrate conflicts in human society based on a common value (Figure 1) would reduce the existential risk. This assumption is consistent with Torres' The Friendly Supersingleton Hypothesis [3]. For stable and continuous operation, the AI system in this paper was constructed as an autonomously distributed system, which has concurrency, scalability, and fault-tolerance. Many issues remain to be solved [17], but this technology is feasible. From this aspect, my method differs from the Friendly Supersingleton of Torres, which is based on future technology. It is desirable for the final form of our proposed AI system to be almost autonomous and worldwide, but part of that system can begin as a conventional AI system with the help of human operators. However, a new issue will then arise regarding executing arbitrations that are consistent with a common value, while avoiding arbitrary influences of human operators.

Finally, various possibilities for applying superintelligence to reduce existential risks caused by various non-AI factors, such as climate change, have been discussed before [31]. With regard to AI itself, discussions have mainly focused on the increase in risks they might pose. An approach using advanced AIs to reduce the existing risks that increase with the progress of AIs has not been sufficiently investigated, either in this paper or by Torres [3]. However, effectively using the power of superintelligence or more elementary AI to construct future governance will create previously unknown possibilities for the future of humanity.

Funding: This research received no external funding.

Acknowledgments: The author thanks Koichi Takahashi, Hiroshi Nakagawa, Hirotaka Osawa, Kentaro Fukuchi, Satoshi Hase, Dohjin Miyamoto, Ichiro Hasegawa, Masahiko Osawa, Ayako Fukawa, and Rzepka Rafal for their helpful comments and feedback.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bostrom, N. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *J. Evol. Technol.* **2001**, *9*, 1.
2. Tinnirello, M. Offensive Realism and the Insecure Structure of the International System: Artificial Intelligence and Global Hegemony. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
3. Torres, P. Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.
4. Sotos, J.G. Biotechnology and the Lifetime of Technical Civilizations. Available online: https://arxiv.org/abs/1709.01149 (accessed on 18 June 2019).
5. Cave, S.; ÓhÉigeartaigh, S.S. An AI Race for Strategic Advantage: Rhetoric and Risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 1–3 February 2018.
6. General AI Challenge. Available online: https://www.general-ai-challenge.org/solving-the-ai-race-results/ (accessed on 9 June 2019).

7. Vincent, M.; Bostrom, N. Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V., Ed.; Springer: Berlin, Germany, 2014.
8. Omohundro, S.M. The Nature of Self-Improving Artificial Intelligence, Presented at the Singularity Summit. Available online: https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf (accessed on 9 June 2019).
9. Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* **2012**, *22*, 71–85. [CrossRef]
10. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
11. Shanahan, M. *The Technological Singularity*; The MIT Press: Cambridge, MA, USA, 2015.
12. Yampolskiy, R.V. Taxonomy of Pathways to Dangerous Artificial Intelligence. In Proceedings of the Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
13. Francois-Lavet, V.; Henderson, P.; Islam, R.; Bellemare, M.G.; Pineau, P. An Introduction to Deep Reinforcement Learning. Available online: https://arxiv.org/abs/1811.12560 (accessed on 9 June 2019).
14. Lem, S. The Invincible (Polish: Niezwyciężony). Available online: https://en.wikipedia.org/wiki/The_Invincible (accessed on 22 June 2019).
15. Hogan, P.J. *Code of the Lifemaker*; Spectrum Literary Agency: New York, NY, USA, 1983.
16. Ahmed, W.; Wu, Y.W. A survey on reliability in distributed systems. *J. Comput. Syst. Sci.* **2013**, *79*, 1243–1255. [CrossRef]
17. Torreño, A.; Onaindia, E.; Komenda, A.; Štolba, M. Cooperative Multi-Agent Planning: A Survey. *ACM Comput. Surv.* **2017**, *50*, 84. [CrossRef]
18. Rao, A.S.; Georgeff, M.P. Modeling rational agents within a BDI-architecture. In Proceedings of the 2nd International Conference Principles of Knowledge Representation and Reasoning, Cambridge, MA, USA, 22–25 April 1991; pp. 473–484.
19. Lynch, N.; Gilbert, S. Conjecture and the feasibility of consistent, available, partition-tolerant web services. *ACM SIGACT News* **2002**, *33*, 51–59.
20. Irving, W. *Rip Van Winkle*; Creative Co.: Boston, MA, USA, 1994.
21. Chaitin, G. *Proving Darwin: Making Biology Mathematical*; Pantheon Books: New York, NY, USA, 2012.
22. Boyce, M.S. Population Viability Analysis. *Annu. Rev. Ecol. Syst.* **1992**, *23*, 481–506. [CrossRef]
23. Nowak, M.A. Five Rules for the Evolution of Cooperation. *Science* **2006**, *314*, 1560–1563. [CrossRef] [PubMed]
24. Beck, K.G.; Zimmerman, K.; Schardt, J.D.; Stone, J.; Lukens, R.R.; Reichard, S.; Randall, J.; Cangelosi, A.A.; Cooper, D.; Thompson, J.P. Invasive Species Defined in a Policy Context: Recommendations from the Federal Invasive Species Advisory Committee. *Invasive Plant Sci. Manag.* **2008**, *1*, 414–421. [CrossRef]
25. Kobayashi, M. *What Is Genome Editing?—The Impact of 'CRISPR'*; Kodansha: Tokyo, Japan, 2016. (In Japanese)
26. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Affective Computing. In *Ethically Aligned Design*, 1st ed.; From Principles to Practice; IEEE Standards Association: Piscataway, NJ, USA, 2019. Available online: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e_affective_computing.pdf (accessed on 9 June 2019).
27. Harari, Y.N. *Sapiens: A Brief History of Humankind*; Harper: New York, NY, USA, 2015.
28. Earle, T.C.; Cvetkovich, G. *Social Trust: Toward a Cosmopolitan Society*; Praeger: Westport, CT, USA, 1995.
29. Mill, J.S. *Utilitarianism*, 1st ed.; Parker, Son & Bourn, West Strand: London, UK, 2015.
30. Cuppen, E. Diversity and constructive conflict in stakeholder dialogue: Considerations for design and methods. *Policy Sci.* **2012**, *45*, 23–46. [CrossRef]
31. Bostrom, N. Strategic Implications of Openness in AI Development. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; Taylor & Francis Group: Boca Raton, FL, USA, 2018.

big data and
cognitive computing

Article

AI Governance and the Policymaking Process: Key Considerations for Reducing AI Risk

Brandon Perry [1,*] and Risto Uuk [2]

[1] Independent Researcher, Berkeley, CA 94709, USA
[2] Effective Altruism Estonia, Tallinn 12618, Estonia; ristouuk@gmail.com
* Correspondence: brandonperryofficial@gmail.com

Received: 5 April 2019; Accepted: 2 May 2019; Published: 8 May 2019

Abstract: This essay argues that a new subfield of AI governance should be explored that examines the policy-making process and its implications for AI governance. A growing number of researchers have begun working on the question of how to mitigate the catastrophic risks of transformative artificial intelligence, including what policies states should adopt. However, this essay identifies a preceding, meta-level problem of how the space of possible policies is affected by the politics and administrative mechanisms of how those policies are created and implemented. This creates a new set of key considerations for the field of AI governance and should influence the action of future policymakers. This essay examines some of the theories of the policymaking process, how they compare to current work in AI governance, and their implications for the field at large and ends by identifying areas of future research.

Keywords: policymaking process; AI risk; typologies of AI policy; AI governance

1. Introduction

Artificial intelligence, especially artificial general intelligence (AGI), has the ability to dramatically impact the future of humanity [1]. Notable researchers, such as Bostrom (2014), have expressed concern that advanced forms of artificial intelligence, if not aligned to humans values and wellbeing, could be potentially disastrous and pose an existential threat to our civilization [2]. The two main branches of research on risk from advanced AI are AI safety, which seeks to ensure that advanced AI is engineered in such a way that it will not pose a threat; and AI governance, which focuses on political and social dynamics (AI macrostrategy) and forecasting timelines for AI development [3]. Issues that AI governance looks at include arms race dynamics, social and economic inequality, public perceptions, issues in surveillance, and more.

There has been a modest amount of work on developing policy solutions to AI risk, with a recent literature review by Baum (2017) [4] and Everitt (2016) [5] covering most of it. Some authors have focused on the development of AGI, with proposed solutions ranging from Joy (2000) [6] who calls for a complete moratorium on AGI research, to Hibbard (2002) [7] and Hughes (2007) [8], who advocate for regulatory regimes to prevent the emergence of harmful AGI, to McGinnis (2010), who advocates for the US to steeply accelerate friendly AGI research [9]. Everitt et al. (2017) [5] suggests that there should be an increase in AI safety funding. Scherer (2016) [10], however, at least in the context of narrow AI, argues that tort law and the existing legal structures, along with the concentration of AI R&D in large visible corporations like Google, will provide some incentives for the safe development of AI. Guihot et al. (2017) [11] also notes that attempts to future-proof laws tend to fail, and pre-emptive bans and regulation tend to hurt the long-term health of the field, instead arguing for a soft-law approach. Other authors have focused on the community of researchers, with Baum (2017) [12] promoting a social psychology approach to promote community self-regulation and activism, and Yampolskiy and Fox (2013) [13] advocating for review boards at universities and other research organizations.

Some authors have advocated for an international approach to resolving AI risk. Erdelyi and Goldsmith (2018) [14] advocated for an international soft-law regime that would serve as a "international forum for discussion and engage in international standard setting activities". Erdelyi and Goldsmith's proposal, however, is not targeted towards AGI risk, although they could scale up to AGI. Wilson (2013) [15] and Bostrom (2014) [2], on the other hand, call for some form of international agreement or control on AGI R&D, with the former advocating specifically for a treaty.

These approaches are necessary given some of the risks, including states pursuing AGI for unprecedented military and economic strength with destabilizing effects (Shulman 2009) [16], and the concentration of wealth and political influence in large corporations (Goertzel 2017) [17]. Questions regarding whether or not AGI R&D should be open sourced or not have been explored by Goertzel (2017) [17] and Bostrom (2017) [18]. Shulman (2009) [16] and Dewey (2015) [19] follow a different approach and advocate for a global surveillance regime to monitor for rogue AGI projects, with Goertzel (2012) [20] suggesting that a limited form of AGI could do this.

As far as current and future research goes, the Future of Humanity Institute has developed an extensive research agenda [3] for AI governance, with three main research areas: Technical landscape, which seeks to understand what artificial intelligence can do and its limits; AI politics, which looks at the political dynamics between firms, governments, publics, etc.; and ideal governance, which looks at possible ways and arrangements for stakeholders to cooperate. This research agenda highlights key issues such as security challenges, international political dynamics and distribution of wealth, and arms race dynamics. Other researchers have published reports dealing with issues such as dual use, similarity, and possible interactions with the cybersecurity community [21] the role and limits of principles for AI ethics [22], justice and equity [23], and AGI R&D community norms [5].

Thus far, much of the literature on AI risk has discussed policy issues, but few studies have talked about how policies are made or how the dynamics of the policymaking process affect their work. Calo (2017) [23] touches upon the problem, noting that there is a lack of institutional expertise, policy tools, and flawed mental models of what AI is, which plague governments' abilities to regulate AI. Scherer (2016) [10] cites certain aspects of the technology itself, such as its ability to be created without special equipment, as a hindrance to the ability to regulate it. Everitt et al. (2017) [5] also briefly discusses policy and political dynamics in the context of AGI researchers, suggesting that AGI researchers should work with other organizations to mitigate the negative dynamics of framing AGI development as an arms race [24]. Finally, the Future of Humanity Institute's research agenda for AI governance [3] touches on policymaking in a few ways, noting that public opinion can have major impacts on technology policy and governance schemes can be subject to mission drift and asking how to facilitate the transition from the present state of affairs to our ideal vision for the future.

This paper continues along the lines of facilitating the transition from the present state to "our ideal vision" by exploring the missing discussion on the role of policymaking in AI governance. Research thus far has largely focused on what problems are out there and what should be done to fix them. However, this paper does not only argue that proposal implementation that takes into account the features of the 'policymaking cycle' may be vital to success in reducing AI risk but that this model actually has massive implications for the research field as a whole. Proposals will be much more effective if they are informed by an understanding of the political and administrative considerations of consensus-building and implementation and could make the difference between making an impact or none at all.

The goal of this paper is to attempt to create a clearer launching point for discussions on the key considerations of the policymaking process for AI governance and the political considerations underpinning policy solutions for AI risk. The policymaking process includes: Problem identification/agenda setting, policy formulation, policy adoption, implementation, and evaluation. Each step of the policymaking process will have different aspects that are critical for the creation of public policies that are able to effectively reduce AI risk. Each section covers a brief overview of the literature, assesses its implications for the greater AI governance field, and identifies different points where further research is needed. The papers we selected are the primary sources of these different theories of the policymaking process.

The first section maps out and defines terms in the field of AI governance, to give readers a better understanding of how our paper contributes to the way AI governance is approached. We also created a typology for AI risk policies, to provide an understanding as to how AI governance has implications in a diverse range of policy communities and how that interplays with strategic considerations. The next section goes through each step of the policymaking cycle, with a basic overview of some of the literature and discussing its implications for AI governance. It should be noted that the literature covered in each field is not extensive, and further research may be necessary. The last sections cover some of the key implications and limitations.

2. Terms and Definitions

On a broad level, the question of mitigating AI risk, or risks that stem from the development and use of artificial intelligence (such as global catastrophic risks from misaligned AI or military instability from adopting new types of weapons), is broken down into AI technical safety and AI governance. AI technical safety focuses on solving computer science problems around issues like misalignment and the control problem for AGI [2]. AI governance, on the other hand, studies how humanity can best navigate the transition to advanced AI systems [3]. This would include the political, military, economic, governance, and ethical considerations and aspects of the problem that advanced AI has on society.

AI governance can be further broken down into other components, namely the technical landscape (how technical developments depends on inputs and constraints and affects rates or domains of capability improvement), ideal governance (what would we do ideally if we could cooperate), and AI politics (how AI will affect domestic politics, political economy, international relations, etc.) [3]. From these research areas, the problems and solutions necessary to discuss AI policy can be defined. This paper, however, refers to this as AI risk policy to differentiate policies intended to reduce catastrophic risk to society versus policies that apply to AI in any other circumstances.

Policies, however, must be implemented into the legal statutes of government in order to work. Flynn (2017) [25], in the blog post that defines 'AI strategy' [3], also defines 'AI policy implementation', which is carrying out the activities necessary to safely navigate the transition to advanced AI systems. This definition implies it is action-oriented work done in government, policy, lobbying, funding, etc. As mentioned in the endnotes of Flynn (2017), however, there is an implicit gap between AI strategy (governance) research and policy implementation, with no AI policy research that identifies mechanisms for actualizing change.

However, there is another gap that this paper intends to address, which is that the processes that create and implement policies (the policymaking process) often either distort the original policy, fall short of, or even work counter to the intended outcome, or render certain policy options unactionable. Similarly, The AI governance: A Research Agenda report has neither this consideration nor a definition of policy implementation. This paper intends to put forth a definition of AI policymaking strategy to fill this gap, which is defined as:

AI Policymaking Strategy: A research field that analyzes the policymaking process and draws implications for policy design, advocacy, organizational strategy, and AI governance as a whole.

This goes further than the concern listed in the endnotes and also develops an upstream approach to AI governance, where work in implementation in turn feeds back and can provide new insights to AI governance research.

AI policymaking strategy would fit under the definition of AI governance and would be its own subfield in the same way technical landscape is and would help to clarify questions and considerations in the other subfields. AI politics and ideal governance seem to ask questions about what risks humanity faces and what it ought to do about them, approaching the world as if from above and making corrections, whereas policymaking strategy asks questions about how and what can be done, given both present and future circumstances, and the methods to do so at hand. They approach the world as agents who individually influence the trajectory of the world. These two groups, when they work together, should ideally converge on a policy program that both works and is pragmatic—constituting of policies that both aim at the correct goals and can actually get there.

An example of this would be the proposed solution by Goertzel (2012) [20] of creating a surveillance artificial narrow intelligence that monitors the world to prevent the development of superintelligence. Let us say that Policy X is written to do this. However, Policy X, like all other policies, is not simply just a solution to the problem but a set of intended actions and procedures taken by the government that must first be passed by government [26]. This begs three questions: Can this policy realistically be implemented by government? How do policymakers ensure that Policy X results in the intended outputs and outcomes? And how can policymakers create policy and advocacy strategies to increase the chances of both of these happening? For example, while Policy X is intended to install a surveillance apparatus to prevent superintelligence, would Policy X still have that output and outcome after going through the legislature and executive branch? Is there a chance over time that it would result in mission creep? Policymakers can also develop strategies to ensure that Policy X has its intended outcomes, such as oversight mechanisms within the policy itself. Policymakers can go a step further and ask how the policymaking process itself creates implications for the AI governance field. For example, are there restrictions within the policymaking process that impact timelines for reducing risk, such as how fast governments can act or create new laws? Could some form of upstream innovation be acheived where the policymaking process inspires or generates new ideas for AI governance [27]?

3. Typologies of AI Policy

Before this paper can delve into the policymaking process, AI policy needs to be further refined to understand what kind of policies are being made. The point of this section is to show that AI risk policies are not monolithic, but rather there are multiple approaches to help achieve the same goal, and each set of these policies is going to have with it a different set of political difficulties. It also begs the question in terms of AI governance as a whole as to which sets of policies should be implemented and when, and which policies should be considered relevant to AI risk. In the same way that Bostrom (2014) [2] argues that there may be a preferred order of technological development, there is a similar analog with AI risk policies where there is a strategic order to policies that should be attempted to be implemented, whether it is because their political-capital cost is lower, the cost of failure is lower, or because it helps with future efforts to implement policies (such as the creation of an advisory body).

A typology of AI policies already has some previous explorative work to build on. Brundage (2016) [28] proposed the idea of De Facto AI policies. These are policies that already exist and are relevant to AI. These are further broken down into direct, indirect, and relevant policies. Direct policies are policies that specifically target AI, such as regulations on self-driving cars. Indirect policies are policies that do not specifically target AI but generally impact the development and diffusion of technologies (including AI), such as intellectual property laws and tort law. Relevant policies do not immediately impact AI but are still worth considering because of their impact, such as education policy or the use of electronic medical records.

Brundage (2016) [27] in this paper, however, does not talk about AI risk policy but rather existing policies around AI as a whole. However, the classification used in this paper is useful overall and can be extended into AI risk policy. Instead of whether or not it directly or indirectly affects AI, AI risk policy can be classified into whether or not it directly or indirectly aims at reducing AI risk. Direct AI risk policies would explicitly govern the use, development, deployment, etc. of AI to reduce risk. Examples of direct AI risk policy could include funding for AI safety research, rules for the development of AGI, international agreements on AI, etc. Indirect AI risk policies would either affect AI but not explicitly govern it or address consequences of the use of advanced AI systems. This could include both policies that affect AI and those that are AI-agnostic. For example, a policy that puts in place stronger protections for privacy in general would reduce the amount of training data available, and thus the speed of AI development, and could be considered an indirect approach. An AI-agnostic policy, for example, would be basic minimum income to address technological unemployment, which could be considered a risk if it leads to societal destabilization. AI risk relevant policies would affect neither AI nor the consequences of it but would rather make it easier for sound AI risk policies to be developed and implemented, such as changing the rules and procedures of government itself to alleviate the pacing problem.

There is another layer of classification that should be applied to AI risk policy based on Lowi's Typology [29]. Lowi categorizes policies into regulatory, distributive, redistributive, and constituency categories. Regulatory policies regulate one's behavior, restricting or incentivizing certain actions, such as the mandating of seat belts in cars. Distributive policies are policies that take money from the general treasury and use them for a specific project that directly benefits one group, such as a dam or research grants. Redistributive policies are those which fundamentally alter the distribution of wealth and resources in the whole of society, such as tax and welfare policies. Constituency policies are those that alter the composition and the rules and regulations of government, such as creating a new executive agency.

Each one of these typologies has with it a certain set of political conditions, as they impact people, businesses, and members of government differently. For example, both basic minimum income and the creation of AI safety standards are policies that are intended to reduce existential risk. However, both of these policies will have a different set of political pressures. Basic minimum income is a redistributive policy, which would move substantial amounts of wealth between classes of society. This would mean that it would likely become a nationwide controversial issue with two opposing camps based largely on who benefits and who loses. By contrast, AI safety standards are a regulatory policy, and while there would be two groups opposed to each other on the issue (unless it comes in the form of voluntary self-regulation by the industry), the political factors around it would look different. Regulatory policies are not usually salient or popular to the general public, and thus, the political battle would be largely limited to regulators, experts, and the business class. This typology will help us to understand how the different policies will be treated in the policymaking process. In other words, policy creates politics. Further work on developing this might be useful for understanding the likelihood of policies being adopted and could shift strategies for which policies to pursue.

4. The Policymaking Cycle

4.1. Problem Identification, Agenda Setting, and Policy Formulation

The first few steps of the policymaking process: Problem identification, agenda setting, and policy formulation, are usually tied together [30], including in a so-called 'multiple streams framework'. The multiple streams framework attempts to explain how policies reach the agenda when policy entrepreneurs are able to couple the policy, politics, and problems streams to open up a policy window, the opportune time when all the conditions are right to get a policy on the agenda [31].

4.1.1. Problem Stream

There are many problems in society. However, the public does not seek government intervention for many of these problems. There are some basic requirements for an issue in society to become a policy problem, which is that it is something that the public finds to be intolerable, government can do something about, and is generally seen as a legitimate area for government to work on [30]. Policy problems can also arise when there are two or more identifiable groups who enter into conflict in a policy arena for resources or positions of power [32].

The first condition for an issue to be considered a policy problem is that it is something that the public or a group finds to be intolerable. Indicators such as statistics can help to identify a problem. These can be used objectively, for understanding conditions in society, or politically, when they are used to justify a political position: for example, using gun violence statistics as an argument for gun control. What is considered an issue over time changes because of the evolution of society. Changes in values, distribution of resources, technology, etc. will change what issues are considered in society [30]. In AI governance, identifiers such as the rate of technological progress or the proliferation of autonomous weapons could be used as examples. Creating a list of politically salient identifiers or metrics could be potentially useful for creating long-term strategies and goals.

How the issue is framed is very important for whether or not it will be considered a policy problem [30]. Is mandating seatbelts in cars beneficial for public safety? Or is it paternalistic? Are these problems legitimate for government to handle? The framing of a problem can have an overwhelming impact on whether or not it is considered a problem appropriate for government to even formulate policy on. It can also impact the content of the policy. Whether you define access to transportation for handicapped people as a transportation problem or a civil rights issue determines whether the acceptable solution involves buying special needs vans, or costly upgrades to buses and subways to ensure equal access. Framing can also raise the priority of a policy problem by, for example, calling it a crisis and raising a sense of urgency.

The question of framing is also incredibly important for AI governance. For example, would autonomous weapons make war more humane by removing humans? Or will it distance ourselves from the violence and make us more willing to use them? The AI governance community needs to think about how these issues ought to be framed, and the consequences of doing so.

In order for an issue to be a part of the system agenda, or what the public or specific communities are discussing, there must be a focusing event. Focusing events are specific events that draw attention to a problem in society and the reasons behind it. The Sandy Hook school shooting, for example, is a focusing event that drew attention to America's gun laws. Moreover, events that occur outside of sector-specific focusing events [31], or past policies on these issues, can have a large impact, especially on the types of solutions used. For AI governance, "Sputnik moments" such as AlphaGo beating Lee Sedol would be an example that drew considerable media attention and generated much discussion about the future of AI, especially in China [33].

Understanding how to exploit these events for the AI governance agenda will be key to generating support and getting policies on the agenda. It is also important to stay on top of these events to understand the direction society is heading in—and to pre-empt or avert less productive or dangerous framings that might feed into arms races [31]. For example, Yampolskiy (2018) details a list of past failures by AI-enabled products [34]. How could work like this be used to influence the problem-setting? Could other AI risk researchers expand on it and build that work into a more thorough project to be used to draw attention to AI risk? Or, could attempts such as this backfire and cause pre-emptive stigmatization or ineffective policies?

4.1.2. Politics Stream

The politics stream is the combined factors of the national mood or public opinion, campaign groups, and administrative/legislative change. Decision-makers in government keep tabs on the swaying opinions of the masses and interest groups and act in a way that promotes themselves favorably, changing items on the agenda to stay relevant and popular, and to obscure unpopular policy stances. Changes in administration, especially when there is a major shift in the ideological composition of the institution, have a strong impact on what is included or not included on the agenda [31].

In AI governance, and for people involved in advocating and implementing policies, maintaining a key eye on domestic and international politics will be key. Knowing when and what kind of policy to advocate for, and to whom, is crucial not only to saving time and energy, but also for legitimacy. Trying to sell a nationalistic administration on greater UN involvement will probably not help someone with furthering their policy proposals and may even damage their (and their coalition's) political capital and cause. However, other forms of cooperation, such as bilateral cooperation for reducing the risk of accidents [35], may be more promising.

AI governance researchers will need to consider how the political landscape should shape their recommendations or policy proposals. Not only would it determine if their recommendations would ever get considered, but if it was implemented, how would it affect the national mood? Would the next administration simply walk it back? How would other interest groups react and impact the long-term ability to reduce risk? If administration changes result in a flip-flop of ideology, what does that mean for AI risk policies associated with the past administration? Could an AI risk policy group maintain influence throughout changing administrations? All of these have implications on our ability to reduce AI risk, and this means that the policymaking strategy will not only have to be robust but also flexible enough to survive changing political conditions.

4.1.3. Policy Stream

The policy stream, which is in essence the policy formulation aspect of the policy cycle, is the "soup" of ideas that are generated by policymakers [35] when deciding what to do about a problem. Different policy networks create policies differently, with different levels of innovativeness and speed [35]. Understanding these differences and examining their implications for the AI governance field might be useful to understand its long-term impact and the specific strategic routes it should take. In other words, how should the AI governance research field itself be organized in a way that promotes useful and relevant solutions?

Despite the staggering number of policy proposals coming out, only a handful will ever be accepted. These policies compete with one another and are selected on a set of criteria, which include technical feasibility, value compatibility [35], budgetary and political costs, and public acceptance. Policies that work will also be technically sound, with no major loopholes, and a clear rationale for how its provisions would lead to actually achieving the policy objectives [30]. This actually creates some key considerations for the field. It means that many ideas are either functionally useless due to their political limitations, unlikely to be adopted in the face of easier or less politically costly options, do not have viable policy mechanisms to achieve their goal, or are otherwise intractable prospects for government. Even if all of the above conditions are resolved, loopholes and unintended consequences may neuter the policy or make conditions worse. This vastly reduces the space of possible solutions. Further, even though the ability for policy implementation or values might change over time, it is still a matter of how much and when. This begs the question: What problems can be solved when, how, and by whom? What does that mean for the large picture strategic approach?

Where should our policies originate from? While there are a bunch of policy ideas out there, only a few are ever seriously considered for adoption. Sources of these policies include (in the United States Federal Government, for example) the President along with the Executive Office of the President, Congressional leaders, government agencies (mostly small incremental changes and adjustments), temporary organizations or 'adhocracies' that serve to investigate specific topics, and interest groups whose topical expertise and political power can sometimes make them de facto policymakers. Each of these areas have differing levels of legitimacy, influence, and degree to which they can make policy changes. A question to consider is not only where in the policy network AI risk policymakers should focus on making these policies, but where they can best advocate for the creation of additional bodies like adhocracies to create additional policies, and what implications that has for the field at large.

With regard to the policy formulation phase of policymaking, a continuum of political environments has been created such that on one extreme, there are policies with publics and on the other, there are policies without publics [36]. When policies are formulated, it is important to consider political environments relevant to the issue. The term "publics" refers to groups who have more than a passing interest in an issue or are actively involved in it. It appears that AI risks are issues where there are limited incentives for publics to form because of problems being remote, costly, or even abstract and uncertain. What does this mean for the AI safety community? How can interest groups be created most effectively? How can these issues be best expressed so that they do not seem so remote, abstract, or uncertain?

4.1.4. Policy Windows and Policy Entrepreneurs

This framework assumes that policy decision-makers, the legislators and bureaucrats in government exist in a state of ambiguity, where they do not have a clear set of preferences, and each set of circumstances can be seen in more than one way. This cannot be resolved with more information, as it is not an issue of ignorance. The example that Zahariadis (2007) gives is that "more information can tell us how AIDS is spread, but it still will not tell us whether AIDS is a health, educational, political, or moral issue [31]".

Overall, the multiple streams framework describes government organizations as "organized anarchies" where institutional problems run rampant, there are often unclear or underdefined goals, overlapping jurisdictions, and a host of other problems that mean that decision-makers have to ration their time between problems and do not have enough time to create a clear set of preferences, make good use of information, or take the time to comprehend the problem for sound decisions on policies. In essence, decision-makers are not rational decision-makers by any stretch. Instead, it depends on the ability of policy entrepreneurs to couple the three streams and manipulate the decision-maker into achieving their intended policy goals [31].

Policy entrepreneurs, who are the policymakers, advocates, interest groups, etc. who push to make specific legislative changes in their areas, only have a short window of time to have their proposals added to the formal agenda. It is when the right political environment, a timely problem, and a potentially acceptable solution all meet together with a policy entrepreneur who can manipulate the situation to their advantage. Because decision-makers exist in a state of ambiguity, policy entrepreneurs are able to manipulate their interpretation of their information to provide meaning, identity, and clarity.

Policy entrepreneurs use different tools and tactics to manipulate the way decision-makers process information and exploit their behavioral biases. Framing tactics, for example, can be used to present a policy option as a loss to the status quo, not taking note of the degree of loss it creates, exploiting decision-makers who are loss-averse, and may push them towards more extreme options like going to war to make up for those small losses [31].

The manipulation of emotions through symbols and the identity or social status of a decision-maker can also pressure them to make certain choices; policies around flag-burning are a great example of this. Because decision-makers are under a great deal of stress and are time-constrained, the strategic ordering of decisions, or 'salami tactics', creates agreement in steps by reducing the total perceived risk of a policy [31]. The manipulation of symbols in the way that artificial intelligence is being framed today has already occured. At first, anti-autonomous weapons advocates were describing 'armed quadcopters' as a serious problem with little media attention [37]. These were rebranded as 'slaughterbots' and a short-film was released with substantial media attention. However, what sort of long-run impact will this have on the field? While giving policymakers straight facts and solutions seems appealing, AI risk policymakers have to consider that it is impractical in reality and may have to accept the inevitability, to policy success, of tactics like framing. Which begs the question, which tactics should they use and how? Questions like these must be considered.

All of this strongly requires an appropriate consideration. Consider, if there are some problems that can only be resolved through state action (such as an arms race), that means that it is dependent on the policymaking process, and thus, these solutions can only be passed when policy windows open. Therefore, how many of these opportunities do AI risk policymakers get? Or, how many chances do they get to implement AI risk policies? These windows only open every once in a while, and they are often in fragile conditions. For example, Bill Clinton's campaign in 1992 aimed to reform the healthcare system and made it a campaign priority, but his administration's failure to pass the bill closed the window [31]. In other words, what impact does this have on AI governance and policy implementation timelines and what does that mean for the field as a whole?

However, in order for a policy entrepreneur to manipulate decision-makers, they must have access to them, which is highly dependent on both the legitimacy of their issue but also for the legitimacy of the group itself and their interest. One of the ways that policy entrepreneurs increase their own influence is to create new decision-points that they can exploit and to reduce access of other groups [32]. AI risk policymakers and advocates will have to find some way to gain access to decision-makers. For example, working on near-term or non-existential risk issues with AI might help someone to build the social capital and network that is necessary to work on existential risks issues. This would not only make it easier people in the field to implement their solutions but to also make themselves gatekeepers to the decision-makers, which could help with preventing policies that would increase existential risks (whether from AI or other sources) from getting through. This may be an area that needs further research. Aspects such as a group's access to decision-makers, the advocating group's legitimacy, biases of the institution [38], and a group's ability to mobilize resources will determine what gets added to the agenda, and the AI risk community will need to work on building all of these. AI policymakers will need to develop a strategy for how to get the right people into the right places and how to coordinate between different groups.

Getting on the formal agenda is a competitive process because there are fundamental limits to a decision-maker's time, and because the policy may be perceived to harm the interests of other groups. Opposing groups can use a variety of tactics, such as denying that the problem exists, arguing that it is not a problem for government, or arguing that the solution would have bad societal consequences, to deny it agenda status. Other factors that could deny an issue agenda status include changing societal norms, political changes, or political leaders avoiding having to be confronted by an issue that hurts their interests. Thus, AI policymakers will need to know how to overcome and adapt to these changing situations and other organizations preventing their policies from being adopted.

AI governance and policy experts will need to pay attention to the arguments being used for and against superintelligence, and whether or not this will become a political issue. Baum (2018) notes that superintelligence is particularly vulnerable to what is known as politicized skepticism, skepticism that is not based on an intellectual disagreement about the problem, based on good-faith attempts to understand the arguments, but rather to shut down concerns based out of self-interest (or a conflict of interests). Some major AI companies, and even other academics, have criticized the idea of superintelligence out of what seems to be their own self-interest as opposed to genuine concerns [39]. This would have a devastating impact on AI policy advocates in a similar way that the tobacco industry significantly impacted scientific efforts to study the public health links between tobacco and cancer.

4.2. Policy Adoption

The next stage of the policy cycle is policy adoption, or when decision-makers choose an option that adopts, modifies, or abandons a policy. This does not necessarily take the form of choosing from a buffet of completed pieces of policy, but rather to take further action on a policy alternative that is more preferable and that is more likely to win approval. At this point, after much bargaining and discussion, the policy choice will only be a formality, or there will be continuous discussion and disagreement until there is a formal vote or decision made. This is an important field to analyze for AI policymakers for the obvious implication that they will want their policy proposals being chosen, and so they will need to understand and design strategies to do so. Further, as will be discussed later, when changes do occur, they can often bring with them wider changes in public policy [40], an implication that will need to be taken into account.

The advocacy coalition framework is a theory on policy adoption but also incorporates every other aspect of the policy cycle with it. The theory describes the interactions of two or more 'advocacy coalitions'; groups of people from a multitude of positions who coordinate together to advocate for some belief, or to implement some policy change (potentially over many fields) over an extended period of time [41]. These do not need to be a single, explicitly delineated organizations like the National Rifle Association but could include loosely affiliated groups of organizations and/or individuals, all working towards the same goal. Building and maintaining coalitions will be one of the major tasks that AI policymakers will need to work on, and so, examining this framework will be highly valuable.

What is it that binds a coalition together? All advocacy coalitions share some form of beliefs. However, the advocacy coalition framework uses a hierarchical belief system. The deepest and broadest of these are deep core beliefs, which are normative positions on human nature, hierarchy of value preferences (i.e., should we value liberty over equality?), the role of government, etc. Policy core beliefs are the next stage of the hierarchy, which involves the extension of deep core beliefs into policy areas. Both of these areas are very difficult to change, as they involve fundamental values. This actually creates an issue where, due to differing fundamental and personal values which lead to lack of interaction, different coalitions often see the same information differently, leading to distrust. Each may come to see the other side as "evil", reducing the possibilities of cooperation and compromise [41].

The deeply held convictions of what a policy subsystem ought to look like are called policy core policy preferences and are the source of conflict between advocacy coalitions. They are the salient problems that have been the long-running issues in that area for a time. Policy core policy preferences shape the political landscape, dictating who allies with whom and who the enemies are, and what strategies coalitions take.

The final level of the belief hierarchy are secondary beliefs, belief that cover procedures, rules, and things of this nature. These are very narrow in scope and the easier to change, requiring less evidence and little bargaining to change.

Understanding the values and beliefs of different existing coalitions, groups, and individuals is key to building and maintaining new coalitions for AI policymakers. This brings up a few considerations. Since it is difficult for conflicting coalitions to work together, will AI policymakers have to choose certain coalitions to work with? What are the costs, benefits, and the potential blowback of this? Since some policies related to AI risk are not in a mature policy field (and thus do not have established coalitions), what can be done to shape the field beforehand to their advantage and/or promote cooperation among coalitions that are likely to form? Further, since secondary beliefs are relatively easy to change, what can be changed to help reduce existential risk?

On a macro-level, this AC Framework acts as a cycle. Relatively stable parameters, as mentioned before, exist in the status quo since policy arenas usually come to some equilibrium where one coalition dominates the policy subsystem. Then, policy changes made by an advocacy coalition or an outside event create a fundamental change in the world, whether it is a change in public opinion or in the rules and procedures governing a subsystem, which changes the initial stable parameters, such as a major event like a mass shooting. These lead to a shift in power that allows another coalition to gain influence over the types of policies being adopted. However, especially in the case of controversial legislature, policies that require multiple veto points to pass will create access for multiple coalitions. This means that even a coalition that dominates a subsystem will not have unilateral ability to dictate policies in some situations. Others, however, especially where there are few decision-makers or an exceptionally influential decision-maker, can result in highly monopolized systems. Questions such as how to be resilient to these changes in conditions, how to facilitate changes into conditions that are beneficial to AI policymakers, and how to construct policy subsystems in a way that is conducive to AI policymakers' goals are useful questions to consider.

This theory describes policy adoption on a very broad level, but how do the decision-makers themselves decide which policies to move forward with? Different incentives and restrictions come to play at different levels of policymaking. For example, highly salient and popular issues are more likely to be influenced by popular opinion, whereas obscure technical issues will likely be determined by policy experts in that field. Different factors that affect both individual and group decision-makers also come into play, such as their personal, professional, organizational, and ideological values. For legislators, their political party and their constituency also play an overwhelming role in their decision-making. Understanding and mapping out these factors will be necessary for the successful implementation of AI risk policy.

On top of these factors, decision-makers usually never have the time, expertise, or even care enough to be able to come up with a fully rational approach to deciding most policies. In many cases, legislators will seek out the advice of other legislators and experts and follow their lead. Due to this being a widespread practice, a few key institutions and leaders often have disproportionate power. For those working in AI risk policy, it is necessary to understand these things so that the message they craft for as to why policy change should occur, and whom to specifically target to get widespread adoption from other decision-makers in the policy arena.

4.3. Policy Implementation

Policy implementation is a key step in the policymaking process. It is defined as "whatever is done to carry a law into effect, to apply it to the target population ... and to achieve its goals" [30]. In other words, it is the activity where adopted policies are carried into effect [30]. However, that is not to say that it is a very distinct step that can be clearly distinguished from others. Every implementation action can influence policy problems, resources, and objectives as the process evolves [42]. Policy implementation can influence problem identification, policy adoption, etc.

Two broad factors that have been offered for the success of policy are local capacity and will [42]. In other words, is there enough training, money, and human resources, along with the right attitudes, motivation, and beliefs to make something happen? It is suggested that the former can be influenced much more easily than the latter as more money can be received and consultants can be hired. For AI risk, both questions are relevant: How to increase capacity and how to influence the influencers. With the former, it has been estimated that about $9-$20 million is currently spent on AI risk [43,44]. With the latter, studying the opinion of the public as well as experts might be a useful approach. One survey [45] indicates that only 8% of top-cited authors in AI consider that human-level AI would be extremely bad (existential risk) for humanity. Another survey that is more recent [46] indicates that machine learning researchers think on average (median) that there is a 10% probability that human-level machine intelligence will result in a negative outcome and 5% probability that it will have an extremely bad outcome (existential risk). The general public seems to be generally cautious, with a survey showing 82% of Americans believing that AI/robots should be managed carefully [47].

This part of the policymaking process is very difficult as the literature is generally quite pessimistic about the ability of policies to bring social changes into effect [48]. However, the authors of the cited paper have identified conditions of effective implementation based on successful examples. These conditions are (a) the policy is based on a sound theory of getting the target group to behave in a desired way, (b) policy directives and structures for the target group are unambiguous, (c) the leaders implementing the policies are skillful with regard to management and politics and committed to the goals, (d) policy is supported by organized constituency groups and key legislators as well as courts throughout the implementation process, and (e) the relative priority of policies is not significantly undermined over time by other policies or socioeconomic changes. Additionally [49], having carefully drafted statute that incentivizes behavior changes, provides adequate funds, expresses clearly ranked goals, is an implementation process, and has few veto points is also vital to the success of a policy.

With regard to AI governance, the ambiguity and complexity of the problem creates a major hurdle for effective policies to be developed. These problems are nonlinear, very hard to predict, and may have the traits of wicked problems in the sense that solving one problem can create new problems. Breaking down AI risk policy into multiple domains as discussed in the previous section helps with creating somewhat less ambiguous objectives, such as changing the education system to be more conducive for technological growth. Even then, however, because many of the issues are either complex or have not happened yet, it is difficult to create concrete objectives and policies. AI risk is not like noise pollution, where there is an easily identifiable, manageable, and tractable problem. Further research could help to identify concrete and tractable issues that might lead to a reduction of risk. In addition, when trying to develop and implement policy, AI policymakers will need to keep in mind factors such as to what extent there is support for it in the executive branch, with outside organizations, and how exactly the policy is written and how those change throughout the policymaking cycle.

Another key consideration for successful policy implementation that was identified from the literature is engaging with the community to increase readiness to accept and devote resources to policy-related problems. It has been acknowledged that there are no good evidence-based ways of achieving community buy-in. This is an area that might be useful to study in order to increase the chances of successful reduction of AI risk. There are different stages of community readiness, such as no awareness, denial, and vague awareness to preplanning, preparation, initiation, and stabilization phases [49]. It is important to understand what counts as the community and what phases different subcommunities of AI safety field are in. Earlier, this paper mentioned a survey about AI experts and showed that their readiness with AI risks was low. Other relevant experts, the public, and other types of subcommunities might have different levels of readiness.

It has been suggested that "the more clearly the core components of an intervention program or practice are known and defined, the more readily the program or practice can be implemented successfully" [49]. In other words, policies and steps of implementation of those policies have to be very clearly expressed. What implications does this have for AI risk? Researchers and policymakers should evaluate how clearly core components have been expressed in this field and improve them as necessary.

4.4. Policy Evaluation

The final step in the policymaking cycle is policy evaluation. This includes activities related to determining the impact of the policy, whether it is achieving its goals, whether the rules and procedures it lays out are being followed, and other externalities or unintended consequences [30]. As we have explained before, policy evaluation does not have to occur only at this step. For example, the impact of a policy is estimated already in the early stages. Anderson et al. highlighted different types of policy evaluations in their book but especially considered systematic evaluations of programs. This involves "the specification of goals or objectives; the collection of information and data on program inputs, outputs, and consequences; and their rigorous analysis, preferably through the use of quantitative or statistical techniques" [30]'.

Policy evaluation examines a policy to understand its impacts in multiple ways [30]. First, is the policy affecting the population that it is intending to target? In AI risk policy, this could be anything from large tech companies, to AI researchers, to people affected by technological unemployment. Second, are there populations that are being affected that were not intended? These externalities could be positive or negative. Third, what are the benefits and costs associated with this policy? AI policymakers will want to ensure that their policies actually reduce risk and that the costs are not so astronomical that they become politically infeasible. Finally, what long-term costs and benefits does a policy have? This is especially important for AI risk policy, as decisions now could have a major impact on the long-term risk that AI has. In AI governance and policymaking, research needs to be done on what sort of indicators or metrics are used for the reduction of risk, and for identifying what goals that should be achieved.

If the previous steps in the policymaking process have generated goals that are unclear or diverse, it is very difficult to evaluate the impact of the policy [30]. Different decision-makers can more easily reach a differing conclusion about the results of a program in that case, or may not follow it all [30]. How the goals of an AI risk program are defined is, therefore, very important.

Another key consideration for policy evaluation is how to make sure that the results are objectively measured. Agency and program officials may be wary of possible political consequences of the evaluation process [30]. If it turns out that the program was not useful or even detrimental, this might have consequences to their influence and career. Because of this consideration, they might not be very interested in correct evaluation studies or they may hinder the process in some other way. There are many ways an evaluation of a policy might be ignored or attacked, such as claiming it was poorly done, the data were inadequate, or the findings inconclusive [30]. Thus, it is important that researchers are provided with high-quality and relevant data-sets that are accurate.

There is also the distinction between policy outputs and outcomes [30] to consider. Outputs are tangible actions taken or things produced, such as collecting taxes or building a dam. Outcomes, on the other hand, are the consequences for society, such as lower disposable income or cleaner air quality. Outputs do not always produce the intended outcomes, which is highly evident in areas such as social welfare policy, where policies may unintentionally trap people in poverty. For AI policymakers, it is very important to consider whether their policy outputs will have the intended consequences, and if so, how to correct that policy.

The evaluation of a policy and the political responses to it can result in the termination of it [30]. Assuming that AI risk policymakers do not want their policies to be terminated or altered in a detrimental way, how can they make sure this does not happen? A policy getting altered to be more effective might be a good thing, but termination can bring unpleasant and negative connotations. It might even have negative consequences to the community [30]. What exact consequences might it have politically? Further, it is important to remember that many policymakers' time horizon only goes until the next election, and so, they often seek immediate results, often before the returns come into fruition. While this may not impact all policies, as this mostly applies to salient policies like healthcare and education, AI policymakers should keep this in mind and try to understand how it might impact their work.

5. Conclusions

There are multiple policy options that could be chosen that either directly or indirectly reduce AI risk, or relevant policies that could help with further efforts to reduce AI risk. Because different policy arenas have different political conditions, and the policymaking process itself draws a number of important challenges, this brings up questions as to what policies in what order are chosen, what strategies are used to get these policies passed and implemented by the government, and the larger impact of these choices on AI governance and risk as a whole. This paper argues that a new subfield of AI governance research on AI policymaking strategies should be further investigated to draw implications for how these policies should be designed, advocated for, and how organizations should approach solving this issue.

6. Limitations and Future Research

This paper is intended to be a broad overview and to be a conversation starter for future research into this area. Thus, there is a strong limitation to the depth of research in this paper. However, it is expected that future work will be done to further refine the line of thinking laid out above, along with further in-depth study into the different theories and their applicability to AI risk.

One of the major limitations of this paper is that the stages heuristic presented in this paper has been heavily criticized and is subject to debate about its effectiveness. Sabatier (2007) has criticized it for not being a causal theory, having a strong top-down bias, among other critiques. However, he also notes that there is much up to debate, with some scholars such as Anderson (2010) advocating for it. There are also a number of other theories that were not discussed in this paper, such as Institutional Rational Choice, the punctuated equilibrium framework, the policy diffusion framework, and other lesser-known theories. Future research is expected that will explore which policy frameworks should be focused on in AI risk research.

The other limitation of this paper is that its applicability to the international governance of AI was not discussed. Future research that looks at how much these theories apply to foreign policy and the international governance of AI in general would be useful. If these theories have a very limited or no impact on the international governance of AI, then figuring out how much work can be done to reduce AI risk in domestic policy would determine the usefulness of these theories.

Throughout the paper, a number of key considerations have been raised. For convenience, a list of them has been curated below below.

7. Summary

This part of the paper summarizes and lists some of the key questions and considerations brought up in the discussion.

Thesis level consideration:

- How do the politics and administrative mechanisms of policymaking affect how policies to mitigate AI risk are created and implemented?

Considerations from Typologies of Policies:

- Are there AI risk policies that should be implemented first? What are the methods to decide this?
- What types of policies should the AI risk policymakers try to get implemented? Why should those types be prioritized?
- What are the political considerations surrounding different sets of policies, and how does that affect their ability to be implemented?

Considerations from Problem Identification, Agenda Setting, and Policy Formulation:

- Is this issue or policy legitimate?
- Would the policy be supported by the current administration and be able to be maintained through changing administrations?
- Which policies out of different sets of potential solutions are politically feasible?
- Are there less costly alternative policies that AI risk policymakers will have to compete with?
- How does attention to problems by different communities affect AI risk policymakers' actions?
- What types of framing of policy issues are most beneficial? What types are most dangerous?
- Is there a way to determine how framing will determine policy content?
- What focusing events have occurred in the field of AI?
- How can AI risk policymakers utilize focusing events to further policy agendas?
- What effect do other organizations have on reducing the legitimacy of AI risk?
- What can be done to respond to these counter-movements effectively? What kind of responses to objections are most convincing?
- How many policy windows will there be for a particular issue? What does this mean for AI risk policymakers' overall strategy?
- What role should AI risk policy entrepreneurs play in AI governance?
- How and where should AI risk policy entrepreneurs gain access in government?

Considerations from Policy Adoption:

- What policy alternatives are more likely to win approval to improve the odds of success for AI risk reduction?
- What strategies can be used to improve the chances of a preferred policy to be adopted?
- Which groups or individuals could join AI risk coalitions, what criteria are used to decide this, and what costs does them joining the coalition have?
- What role can organizations outside of AI risk play in furthering AI risk policymakers' agenda?

Considerations from Policy Implementation:

- Is this solution technically feasible for governments to implement?

- Are there enough resources, will, and support by leaders and constituency groups to be successful in implementation?
- Is the policy crafted in a way that effectively structures incentives for the target group?
- Is the policy unambiguous? If so, then how will that affect its ability to be implemented?
- Are the goals of the policy in conflict with any other policy or changes in society?
- Are there any veto points in the policy's statutes to prevent effective implementation?
- How will the contents or the political factors surrounding of a policy be affected during implementation?
- Do the relevant communities accept the issue, and are they willing to devote resources to resolve it?

Considerations from Policy Evaluation:

- Are the policy outputs having the intended outcomes?
- What are the consequences of any unintentional outcomes?
- What are the political factors surrounding the metrics that are being used to evaluate the policy?
- Do the political costs or benefits of the policy have an impact on its success?
- If the policy is terminated, will there be any negative political consequences?
- How can AI risk policymakers update the policy? How can they prevent changes by other groups that would be harmful?
- How will the limited time horizons of lawmakers and other groups affect the evaluation of the policy?

Author Contributions: Conceptualization, B.P. and R.U.; methodology, B.P. and R.U.; writing—original draft preparation, B.P. and R.U.; writing—review and editing, B.P. and R.U.; project administration, B.P. and R.U.

Funding: This research received no external funding.

Acknowledgments: The authors thank Matthijs Maas, Seth Baum, Sabrina Kavanaugh, Max Daniel, and the organizers and participants of the AI Safety Camp for useful comments and feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References and Notes

1. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
2. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK; New York, NY, USA, 2014.
3. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute: Oxford, UK, 2018. Available online: https://www.fhi.ox.ac.uk/govaiagenda/ (accessed on 17 December 2018).
4. Baum, S.D. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. Global Catastrophic Risk Institute Working Paper 17-1. 2017. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741 (accessed on 11 November 2019).
5. Everitt, T.; Lea, G.; Hutter, M. AGI Safety Literature Review. *arXiv* **2018**, arXiv:1805.01109.
6. Joy, B. Why the future doesn't need us. *Wired* **2000**, *8*, 238–263. Available online: https://www.wired.com/2000/04/joy-2/ (accessed on 6 January 2019).
7. Hibbard, B. *Super-Intelligent Machines*; Springer: New York, NY, USA, 2002.
8. Hughes, J.J. Global technology regulation and potentially apocalyptic technological threats. In *Nanoethics: The Ethical and Social Implications of Nanotechnology*; Allhoff, F., Ed.; John Wiley: Hoboken, NJ, USA, 2007; pp. 201–214.
9. McGinnis, J.O. Accelerating AI. *Northwest. Univ. Law Rev.* **2010**, *104*, 366–381. Available online: https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1193&context=nulr_online (accessed on 14 March 2019). [CrossRef]
10. Scherer, M.U. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. J. Law Technol.* **2016**, *29*, 354–398. [CrossRef]
11. Guihot, M.; Matthew, A.F.; Suzor, N.P. Nudging robots: Innovative solutions to regulate artificial intelligence. *Vanderbilt J. Entertain. Technol. Law* **2017**, *20*, 385–456.

12. Baum, S.D. On the promotion of safe and socially beneficial artificial intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
13. Yampolskiy, R.; Fox, J. Safety Engineering for Artificial General Intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
14. Erdelyi, O.J.; Goldsmith, J. Regulating Artificial Intelligence: Proposal for a Global Solution. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), New Orleans, LO, USA, 2–3 February 2018. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3263992 (accessed on 6 January 2019).
15. Wilson, G. Minimizing global catastrophic and existential risks from emerging technologies through international law. *Va. Environ. Law J.* **2013**, *31*, 307–364.
16. Shulman, C. Arms control and intelligence explosions. In Proceedings of the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, 2–4 July 2009.
17. Goertzel, B. The Corporatization of AI is a Major Threat to Humanity. h+ Magazine. 2017. Available online: http://hplusmagazine.com/2017/07/21/corporatization-ai-major-threat-humanity/ (accessed on 6 January 2019).
18. Bostrom, N. Strategic Implications of Openness in AI Development. *Glob. Policy* **2017**, *8*, 135–148. [CrossRef]
19. Dewey, D. Long-term strategies for ending existential risk from fast takeoff. In *Risks of Artificial Intelligence*; Müller V.C., Ed.; CRC: Boca Raton, FL, USA, 2015; pp. 243–266.
20. Goertzel, B. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? *J. Conscious. Stud.* **2012**, *19*, 96.
21. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B.; et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Available online: https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/1c6q2kc4v_50335.pdf (accessed on 6 January 2018).
22. Whittlestone, J.; Nyrup, R.; Alexandrova, A.; Cave, S. The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA, 27–28 January 2019.
23. Calo, R. Artificial Intelligence Policy: A Primer and Roadmap. 2017. Available online: https://ssrn.com/abstract=3015350 (accessed on 6 January 2019). It should also be noted that Calo is dismissive of the risk of artificial general intelligence.
24. Cave, S.; ÓhÉigeartaigh, S.S. An AI Race for Strategic Advantage: Rhetoric and Risks. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, HI, USA, 27–28 January 2019. Available online: http://www.aies-conference.com/wp-content/papers/main/AIES_2018_paper_163.pdf (accessed on 14 March 2019).
25. Flynn, C. Personal Thoughts on Careers in AI Policy and Strategy. Effective Altruism Forum. 2017. Available online: https://forum.effectivealtruism.org/posts/RCvetzfDnBNFX7pLH/personal-thoughts-on-careers-in-ai-policy-and-strategy (accessed on 6 January 2019).
26. The specifics issues will depend on the type of government. For example, the types of difficulties would be different in a democracy vs. a dictatorship. This paper however will focus on federal republics.
27. Thank you to Sabrina Kavanagh for suggesting the idea that the policy process could inspire new ideas for AI governance researchers.
28. Brundage, M.; Bryson, J. Smart Policies for Artificial Intelligence. *arXiv* **2016**, arXiv:1608.08196.
29. Lowi, T.J. Four Systems of Policy, Politics, and Choice. *Public Adm. Rev.* **1972**, *32*, 298–310. [CrossRef]
30. Anderson, J.E. *Public Policymaking: An Introduction*, 7th ed.; Cengage Learning: Boston, MA, USA, 2010.
31. Zahariadis, N. The Multiple Streams Framework: Structure, Limitations, Prospects. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P., Eds.; Westview Press: Boulder, CO, USA, 2007.
32. Cobb, R.; Elder, C.D. What is an Issue? What Makes an Issue? In *Participation in American Politics: The Dynamics of Agenda Building*; Johns Hopkins University Press: Baltimore, MD, USA, 1983; pp. 82–93.
33. Allen, G. China's Artificial Intelligence Strategy Poses a Credible Threat to U.S. Tech Leadership. Center for Foreign Affairs Blog. Available online: https://www.cfr.org/blog/chinas-artificial-intelligence-strategy-poses-credible-threat-us-tech-leadership (accessed on 26 February 2019).
34. Yampolskiy, R. Current State of Knowledge on Failures of AI Enabled Products. Report. Consortium for Safer AI. 2018. Available online: https://docs.wixstatic.com/ugd/ace275_0ea60fe9b665439bb0b37d20beb89b6f.pdf (accessed on 6 January 2018).

35. Danzig, R. *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*; Center for New American Security: Washington, DC, USA, 2018. Available online: https://www.cnas.org/publications/reports/technology-roulette (accessed on 24 March 2019).
36. May, P.J. Reconsidering Policy Design: Policies and Publics. *J. Public Policy* **1991**, *11*, 187–206. [CrossRef]
37. Russell, S.; Aguirre, A.; Conn, A.; Tegmark, M. Why You Should Fear "Slaughterbots"—A Response. IEEE Spectrum. 2018. Available online: https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-you-should-fear-slaughterbots-a-response (accessed on 9 January 2019).
38. Yudkowsky, E. Cognitive Biases Potentially Affecting Judgment of Global Risks. In *Global Catastrophic Risks*; Bostrom, N., Ćirković, M.M., Eds.; Oxford University Press: New York, NY, USA, 2008; pp. 91–119.
39. Baum, S.D. Superintelligence Skepticism as a Political Tool. *Information* **2018**, *9*, 209. [CrossRef]
40. James, T.L.; Jones B.D.; Baumgartner, F.R. Punctuated-Equilibrium Theory: Explaining Stability and Change in Public Policymaking. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P.A., Ed.; Westview Press: Boulder, CO, USA, 2007; Chapter 6.
41. Sabatier, P.; Weiblle, C.M. An Advocacy Coalition Framework. In *Theories of the Policy Process*, 2nd ed.; Sabatier, P.A., Ed.; Westview Press: Boulder, CO, USA, 2007; Chapter 7.
42. McLaughlin, M.W. Learning From Experience: Lessons From Policy Implementation. *Educ. Eval. Policy Anal.* **1987**, *9*, 171–178. [CrossRef]
43. Farquhar, S. Changes in Funding in the AI Safety Field. 2017. Available online: https://www.centreforeffectivealtruism.org/blog/changes-in-funding-in-the-ai-safety-field (accessed on 6 January 2019).
44. MacAskill, W. What Are the Most Important Moral Problems Of Our Time? TED Talk. 2018. Available online: https://www.ted.com/talks/will_macaskill_how_can_we_do_the_most_good_for_the_world (accessed on 6 January 2019).
45. Müller, V; Bostrom, N. Future progress in artificial intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*; Müller, V.C., Ed.; Synthese Library; Springer: Berlin, Germany, Forthcoming 2014. Available online: https://nickbostrom.com/papers/survey.pdf (accessed on 6 January 2019).
46. Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B.; Evans, O. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv* **2017**, arXiv:1705.08807.
47. Zhang, B.; Dafoe, A. Artificial Intelligence: American Attitudes and Trends. January 2019. Available online: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf (accessed on 3 January 2019).
48. Sabatier, P.; Mazmanian, D. The Conditions of Effective Implementation: A Guide to Accomplishing Policy Objectives. *Policy Anal.* **1979**, *5*, 481–504. [PubMed]
49. Sabatier, P.; Mazmanian, D. The Implementation of Public Policy: A Framework of Analysis. *Policy Stud. J.* **1980**, *8*, 538–560. [CrossRef]

big data and
cognitive computing

MDPI

Communication

Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence

David Manheim

Independent Researcher, Silver Spring, MD 20910, USA; davidmanheim@gmail.com

Received: 28 February 2019; Accepted: 29 March 2019; Published: 5 April 2019

Abstract: An important challenge for safety in machine learning and artificial intelligence systems is a set of related failures involving specification gaming, reward hacking, fragility to distributional shifts, and Goodhart's or Campbell's law. This paper presents additional failure modes for interactions within multi-agent systems that are closely related. These multi-agent failure modes are more complex, more problematic, and less well understood than the single-agent case, and are also already occurring, largely unnoticed. After motivating the discussion with examples from poker-playing artificial intelligence (AI), the paper explains why these failure modes are in some senses unavoidable. Following this, the paper categorizes failure modes, provides definitions, and cites examples for each of the modes: accidental steering, coordination failures, adversarial misalignment, input spoofing and filtering, and goal co-option or direct hacking. The paper then discusses how extant literature on multi-agent AI fails to address these failure modes, and identifies work which may be useful for the mitigation of these failure modes.

Keywords: multi-agent systems; specification gaming; artificial intelligence safety; Goodhart's Law

MSC: 91E45; 91A06

JEL Classification: C79; D74

1. Background, Motivation and Contribution

When complex systems are optimized by a single agent, the representation of the system and of the goal used for optimization often leads to failures that can be surprising to the agent's designers. These failure modes go by a variety of names, Amodei and Clark called them faulty reward functions [1] but similar failures have been referred to as Goodhart's law [2,3], Campbell's law [4], distributional shift [5], strategic behavior [6], reward hacking [7], Proxyeconomics [8], and other terms.

Examples of these failures in the single-agent case are shown by Victoria Krakovna's extensive list of concrete examples of "generating a solution that literally satisfies the stated objective but fails to solve the problem according to the human designer's intent." [9] Liu et al. suggest that "a complex activity can often be performed in several different ways," [10] but not all these ways should be considered valid. To understand why, Krakovna' s list includes examples of "achieving a goal" by finding and exploiting bugs in a simulation engine to achieve goals [11–13]; by physical manipulation of objects in unanticipated ways, such as moving a table instead of the item on the table [14], or flipping instead of lifting a block [15]; and even by exploiting the problem structure or evaluation, such as returning an empty list as being sorted [16], or deleting the file containing the target output [16].

1.1. Motivation

This forms only a part of the broader set of concerns in AI safety, [5,17–19], but the failure modes are the focus of a significant body of work in AI safety discussed later in the paper. However, as the systems become more capable and more widely used, Danzig and others have noted that this will "increase rather than reduce collateral risks of loss of control." [20] The speed of such systems is almost certainly beyond the point of feasible human control, and as they become more complex, the systems are also individually likely to fail in ways that are harder to understand.

While some progress has been made in the single-agent case, the systems have continued to become more capable, corporations, governments, and other actors have developed and deployed machine learning systems. These are not only largely autonomous, but also interact with each other. This allows a new set of failures, and these are not yet a focus of safety-focused research—but they are critical.

1.2. Contribution

The analogues of the earlier-mentioned classes of failure for multi-agent systems are more complex, potentially harder to mitigate, and unfortunately not the subject of a significant focus among AI safety researchers. In this paper, we introduce a classification of failures that are not yet well-addressed in the literature involving multiple agents. These failures can occur even when system designers do not intend to build conflicting AI or ML systems. The current paper contributes to the literature by outlining how and why these multi-agent failures can occur, and providing an overview of approaches that could be developed for mitigating them. In doing so, the paper will hopefully help spur system designers to explicitly consider these failure modes in designing systems, and urge caution on the part of policymakers.

As a secondary contribution, the link between ongoing work on AI safety and potential work mitigating these multi-agent failures incidentally answers an objection raised by AI risk skeptics that AI safety is "not worth current attention" and that the issues are "premature to worry about" [21]. This paper instead shows how failures due to multi-agent dynamics are critical in the present, as ML and superhuman narrow AI is being widely deployed, even given the (valid) arguments put forward by Yudkowsky [22] and Bostrom [7] for why a singleton AI is a more important source of existential risk.

1.3. Extending Single-Agent Optimization Failures

Systems which are optimized using an imperfect system model have several important failure modes categorized in work by Manheim and Garrabrant [3]. First, imperfect correlates of the goal will be less correlated in the tails of the distribution, as discussed by Lewis [23]. Heavily optimized systems will end up in those regions, and even well-designed metrics do not account for every possible source of variance. Second, there are several context failures [24], where the optimization is well behaved in the training set ("ancestral environment") but fails as optimization pressure is applied. For example, it may drift towards an "edge instantiation" where the system may optimize all the variables that relate to the true goal, but further gain on the metric is found by unexpected means. Alternatively, the optimizer may properly obey constraints in the initial stage, but find some "nearest unblocked strategy" [24] allowing it to circumvent designed limits when given more optimization power. These can all occur in single-agent scenarios.

The types of failure in multi-agent systems presented in this paper can be related to Manheim and Garrabrant's classification of single-agent metric optimization failures . The four single-agent overoptimization failure modes outlined there are:

- Tails Fall Apart, or Regressional inaccuracy, where the relationship between the modeled goal and the true goal is inexact due to noise (for example, measurement error,) so that the bias grows as the system is optimized.
- Extremal Model Insufficiency, where the approximate model omits factors which dominate the system's behavior after optimization.
- Extremal Regime Change, where the model does not include a regime change that occurs under certain (unobserved) conditions that optimization creates.

- Causal Model Failure, where the agent's actions are based on a model which incorrectly represents causal relationships, and the optimization involves interventions that break the causal structure the model implicitly relies on.

Despite the completeness of the above categorization, the way in which these failures occur can differ greatly even when only a single agent is present. In a multi-agent scenario, agents can stumble into or intentionally exploit model overoptimization failures in even more complex ways. Despite this complexity, the different multi-agent failure modes can be understood based on understanding the way in which the implicit or explicit system models used by agents fail.

1.4. Defining Multi-Agent Failures

In this paper, a multi-agent optimization failure is when one (or more) of the agents which can achieve positive outcomes in some scenarios exhibits behaviors that negatively affect its own outcome due to the actions of one or more agents other than itself. This occurs either when the objective function of the agent no longer aligns with the goal, as occurs in the Regressional and both Extremal cases, or when the learned relationship between action(s), the metric(s), and the goal have changed, as in the Causal failure case.

This definition does not require the failure to be due to malicious behavior on the part of any agent, nor does it forbid it. Note also that the definition does not require failure of the system, as in Behzadan and Munir's categorization of adversarial attacks [25], nor does it make any assumptions about type of the agents, such as the type of learning or optimization system used. (The multi-agent cases implicitly preclude agents from being either strongly boxed, as Drexler proposed [26], or oracular, as discussed by Armstrong [27].)

2. Multi-Agent Failures: Context and Categorization

Several relatively straightforward failure modes involving interactions between an agent and a regulator were referred to in Manheim and Garrabrant as adversarial Goodhart [3]. These occur where one AI system opportunistically alters or optimizes the system and uses the expected optimization of a different victim agent to hijack the overall system. For example, "smart market" electrical grids use systems that optimize producer actions and prices with a linear optimization system using known criteria. If power lines or power plants have strategically planned maintenance schedules, an owner can manipulate the resulting prices to its own advantage, as occurred (legally) in the case of Enron [28]. This is possible because the manipulator can plan in the presence of a known optimization regime.

This class of manipulation by an agent frustrating a regulator's goals is an important case, but more complex dynamics can also exist, and Manheim and Garrabrant noted that there are "clearly further dynamics worth exploring." [3] This involves not only multiple heterogenous agents, which Kleinberg and Raghavan suggest an avenue for investigating, but also interaction between those agents [6]. An example of a well-understood multi-agent system, the game of poker, allows clarification of why the complexity is far greater in the interaction case.

2.1. Texas Hold'em and the Complexity of Multi-Agent Dynamics

In many-agent systems, simple interactions can become complex adaptive systems due to agent behavior, as the game of poker shows. Solutions to simplified models of two-player poker predate game theory as a field [29], and for simplified variants, two-player draw poker has a fairly simple optimal strategy [30]. These early, manually computed solutions were made possible both by limiting the complexity of the cards, and more importantly by limiting interaction to a single bet size, with no raising or interaction between the players. In the more general case of heads-up limit Texas Hold'em, significantly more work was needed, given the multiplicity of card combinations, the existence of hidden information, and player interaction, but this multi-stage interactive game is "now essentially weakly solved" [31]. Still, this game involves only two players. In the no-limit version of the game, Brown and Sandholm recently unveiled superhuman AI [32], which restricts the game to "Heads' Up" poker, which involves only two players per game, and still falls far short of a full solution to the game.

The complex adaptive nature of multi-agent systems means that each agent needs model not only model the system itself, but also the actions of the other player(s). The multiplicity of potential outcomes, betting strategies, and different outcomes becomes rapidly infeasible to represent other than heuristically. In limit Texas Hold'em poker, for example, the number of card combinations is immense, but the branching possibilities for betting is the more difficult challenge. In a no-betting game of Hold'em with P players, there are $52!/((52-2P-5)!\cdot 2P\cdot 5!)$ possible situations. This is $2.8\cdot 10^{12}$ hands in the two-player case, $3.3\cdot 10^{15}$ in the three-player case, and growing by a similar factor when expanded to the four-, five-, or six-player case. The probability of winning is the probability that the five cards on the table plus two unknown other cards from the deck are a better hand than any that another player holds. In Texas Hold'em, there are four betting stages, one after each stage of cards is revealed. Billings et al. use a reduced complexity game (limiting betting to three rounds per stage) and find a complexity of $O(10^{18})$ in the two-hand case [33]. That means the two-player, three-round game complexity is comparable in size to a no-betting four-player game, with $4.1\cdot 10^{18}$ card combinations possible.

Unlike a no-betting game, however, a player must consider much more than the simple probability that the hand held is better than those held by other players. That calculation is unmodified during the additional branching due to player choices. The somewhat more difficult issue is that the additional branching requires Bayesian updates to estimate the probable distribution of hand strengths held by other players based on their decisions, which significantly increases the complexity of solving the game. The most critical challenge, however, is that each player bets based on the additional information provided by not only the hidden information provided by their cards, but also based on the betting behavior of other players. Opponent(s) make betting decisions based on non-public information (in Texas Hold'em, hole cards) and strategy for betting requires a meta-update taking advantage of the information the other player reveals by betting. The players must also update based on potential strategic betting by other players, which occurs when a player bets in a way calculated to deceive. To deal with this, poker players need to model not just the cards, but also the strategic decisions of other players. This complex model of strategic decisions must be re-run for all the possible combinations at each decision point to arrive at a conclusion about what other players are doing. Even after this is complete, an advanced poker player, or an effective AI, must then decide not just how likely they are to win, but also how to play strategically, optimizing based on how other players will react to the different choices available.

Behaviors such as bluffing and slow play are based on these dynamics, which become much more complex as the number of rounds of betting and the number of players increases. For example, slow play involves underbidding compared to the strength of your hand. This requires that the players will later be able to raise the stakes, and allows a player to lure others into committing additional money. The complexity of the required modeling of other agents' decision processes grows as a function of the number of choices and stages at which each agent makes a decision. This type of complexity is common in multi-agent systems. In general, however, the problem is much broader in scope than what can be illustrated by a rigidly structured game such as poker.

2.2. Limited Complexity Models versus the Real World

In machine learning systems, the underlying system is approximated by implicitly or explicitly learning a multidimensional transformation between inputs and outputs. This transformation approximates a combination of the relationships between inputs and the underlying system, and between the system state and the outputs. The complexity of the model learned is limited by the computational complexity of the underlying structure, and while the number of possible states for the input is large, it is typically dwarfed by the number of possible states of the system.

The critical feature of machine learning that allows such systems to be successful is that most relationships can be approximated without inspecting every available state. (All models simplify the systems they represent.) The implicit simplification done by machine learning is often quite impressive, picking up on clues present in the input that humans might not notice, but it comes at the cost of having difficult to understand and difficult to interpret implicit models of the system.

Any intelligence, whether machine learning-based, human, or AI, requires similar implicit simplification, since the branching complexity of even a relatively simple game such as Go dwarfs the number of atoms in the universe. Because even moderately complex systems cannot be fully represented, as discussed by Soares [34], the types of optimization failures discussed above are inevitable. The contrapositive to Conant and Ashby's theorem [35] is that if a system is more complex than the model, any attempt to control the system will be imperfect. Learning, whether human or machine, builds approximate models based on observations, or input data. This implies that the behavior of the approximation in regions far from those covered by the training data is more likely to markedly differ from reality. The more systems change over time, the more difficult prediction becomes—and the more optimization is performed on a system, the more it will change. Worsening this problem, the learning that occurs in ML systems fails to account for the embedded agency issues discussed by Demski and Garrabrant [36], and interaction between agents with implicit models of each other and themselves amplifies many of these concerns.

2.3. Failure modes

Because an essential part of multi-agent dynamic system modeling is opponent modeling, the opponent models are a central part of any machine learning model. These opponent models may be implicit in the overall model, or they may be explicitly represented, but they are still models that are approximate. In many cases, opponent behavior is ignored—by implicitly simplifying other agent behavior to noise, or by assuming no adversarial agents exist. Because these models are imperfect, they will be vulnerable to overoptimization failures discussed above.

The list below is conceptually complete, but limited in at least three ways. First, examples given in this list primarily discuss failures that occur between two parties, such as a malicious actor and a victim, or failures induced by multiple individually benign agents. This would exclude strategies where agents manipulate others indirectly, or those where coordinated interaction between agents is used to manipulate the system. It is possible that when more agents are involved, more specific classes of failure will be relevant.

Second, the below list does not include how other factors can compound metric failures. These are critical, but may involve overoptimization, or multiple-agent interaction, only indirectly. For example, O'Neil discusses a class of failure involving the interaction between the system, the inputs, and validation of outputs [37]. These failures occur when a system's metrics are validated in part based on outputs it contributes towards. For example, a system predicting greater crime rates in areas with high minority concentrations leads to more police presence, which in turn leads to a higher rate of crime found. This higher rate of crime in those areas is used to train the model, which leads it to reinforce the earlier unjustified assumption. Such cases are both likely to occur, and especially hard to recognize, when the interaction between multiple systems is complex, and it is unclear whether the system's effects are due in part to its own actions (This class of failure seems particularly likely in systems that are trained via ”self-play,” where failures in the model of the system get reinforced by incorrect feedback on the basis of the models, which is also a case of model insufficiency failure.).

Third and finally, the failure modes exclude cases that do not directly involve metric overoptimizations, such as systems learning unacceptable behavior implicitly due to training data that contains unanticipated biases, or failing to attempt to optimize for social preferences such as fairness. These are again important, but they are more basic failures of system design.

With those caveats, we propose the following classes of multi-agent overoptimization failures. For each, a general definition is provided, followed by one or more toy models that demonstrate the failure mode. Each agent attempts to achieve their goal by optimizing for the metric, but the optimization is performed by different agents without any explicit coordination or a priori knowledge about the other agents. The specifics of the strategies that can be constructed and the structure of the system can be arbitrarily complex, but as explored below, the ways in which these models fail can still be understood generally.

These models are deliberately simplified, but where possible, real-world examples of the failures exhibited in the model are suggested. These examples come from both human systems where parallel dynamics exist, and examples of the failures in extent systems with automated agents. In the toy models, M_i and G_i stands for the metric and goal, respectively, for agent i. The metric is an imperfect proxy for the goal, and will typically be defined in relation to a goal. (The goal itself is often left unspecified, since the model applies to arbitrary systems and agent goals.) In some cases, the failure is non-adversarial, but where relevant, there is a victim agent V and an opponent agent O that attempts to exploit it. Please note that the failures can be shown with examples formulated with game-theoretic notation, but doing so requires more complex specifications of the system and interactions than is possible using the below characterization of the agent goals and the systems.

Failure Mode 1. ***Accidental Steering*** *is when multiple agents alter the systems in ways not anticipated by at least one agent, creating one of the above-mentioned single-party overoptimization failures.*

Remark 1. *This failure mode manifests similarly to the single-agent case and differs only in that agents do not anticipate the actions of other agents. When agents have closely related goals, even if those goals are aligned, it can exacerbate the types of failures that occur in single-agent cases.*

Because the failing agent alone does not (or cannot) trigger the failure, this differs from the single-agent case. The distributional shift can occur due to a combination of actors' otherwise potentially positive influences by either putting the system in an extremal state where the previously learned relationship decays, or triggering a regime change where previously beneficial actions are harmful.

Model. ***1.1—Group Overoptimization.*** *A set of agents each have goals which affect the system in related ways, and metric-goal relationship changes in the extremal region where x>a. As noted above, M_i and G_i stands for the metric and goal, respectively, for agent i. This extremal region is one where single-agent failure modes will occur for some or all agents. Each agent i can influence the metric by an amount α_i, where $\sum \alpha_i > a$, but $\forall \alpha_i < a$. In the extremal subspace where $M_i > a$, the metric reverses direction, making further optimization of the metric harm the agent's goal.*

$$M_i = \begin{cases} G_i, & \text{where } M_i <= a \\ M_i(a) - G_i, & \text{where } M_i > a \end{cases} \tag{1}$$

Remark 2. *In the presence of multiple agents without coordination, manipulation of factors not already being manipulated by other agents is likely to be easier and more rewarding, potentially leading to inadvertent steering due to model inadequacy, as discussed in Manheim and Garrabrant's categorization of single-agent cases [3]. As shown there overoptimization can lead to perverse outcomes, and the failing agent(s) can hurt both their own goals, and in similar ways, can lead to negative impacts on the goals of other agents.*

Model. ***1.2—Catastrophic Threshold Failure.***

$$M_i = x_i \qquad G_i = \begin{cases} a + (\sum_{\forall i} x_i) & \text{where } \sum_{\forall i} x_i <= T \\ a - - - (\sum_{\forall i} x_i) & \text{where } \sum_{\forall i} x_i > T \end{cases} \tag{2}$$

Each agent manipulates their own variable, unaware of the overall impact. Even though the agents are collaborating, because they cannot see other agents' variables, there is no obvious way to limit the combined *impact on the system to stay below the catastrophic threshold T. Because each agent is exploring a different variable, they each are potentially optimizing different parts of the system.*

Remark 3. *This type of catastrophic threshold is commonly discussed in relations to complex adaptive systems, but can occur even in systems where the catastrophic threshold is simple. The case discussed by Michael Eisen involves pricing on Amazon was due to a pair of deterministic linear pricing-setting bots interacting to set the price of an otherwise unremarkable biology book at tens of millions of dollars, showing that runaway dynamics are possible even in the simplest cases [38]. This phenomenon is also expected whenever exceeding some constraint breaks the system, and such constraints are often not identified until a failure occurs.*

Example 1. *This type of coordination failure can occur in situations such as overfishing across multiple regions, where each group catches local fish, which they can see, but at a given threshold across regions the fish population collapses, and recovery is very slow. (In this case, the groups typically are selfish rather than collaborating, making the dynamics even more extreme.)*

Example 2. *Smaldino and McElreath [39] shows this failure mode specifically occurring with statistical methodology in academia, where academics find novel ways to degrade statistical rigor. The more general "Mutable Practices" model presented by Braganza [8], based on part on Smaldino and McElreath, has each agent attempting to both outperform the other agents on a metric as well as fulfill a shared societal goal, allows agents to evolve and find new strategies that combine to subvert a societal goal.*

Failure Mode 2. ***Coordination Failure*** *occurs when multiple agents clash despite having potentially compatible goals.*

Remark 4. *Coordination is an inherently difficult task, and can in general be considered impossible [40]. In practice, coordination is especially difficult when the goals of other agents are incompletely known or not fully understood. Coordination failures such as Yudkowsky's Inadequate equilibria are stable, and coordination to escape from such an equilibrium can be problematic even when agents share goals [41].*

Model. ***2.1—Unintended Resource Contention.*** *A fixed resource R is split between uses R^n by different agents. Each agent has limited funds f_i, and R_i is allocated to agent i for exploitation in proportion to their bid for the resources c_{R_i}. The agents choose amounts to spend on acquiring resources, and then choose amounts s_{n_i} to exploit each resource, resulting in utility $U(s_n, R_n)$. The agent goals are based on the overall exploitation of the resources by all agents.*

$$\begin{aligned} R_i &= \frac{c_{R_i}}{\sum_{\forall i} c_{R_i}} \\ G_i &= \sum_{\forall i} U_{i,n}(s_{n_i}, R_n) \end{aligned} \tag{3}$$

In this case, we see that conflicting instrumental goals that neither side anticipates will cause wasted funds due to contention. The more funds spent on resource capture, which is zero-sum, the less remaining for exploitation, which can be positive-sum. Above nominal spending on resources to capture them from aligned competitor-agents will reduce funds available for exploitation of those resources, even though less resource contention would benefit all agents.

Remark 5. *Preferences and gains from different uses can be homogeneous, so that all agents have no marginal gain from affecting the allocation, funds will be wasted on resource contention. More generally, heterogeneous preferences can lead to contention to control the allocation, with sub-optimal individual outcomes, and heterogeneous abilities can lead to less-capable agents harming their goals by capturing then ineffectively exploiting resources.*

Example 3. *Different forms of scientific research benefit different goals differently. Even if spending in every area benefits everyone, a fixed pool of resources implies that with different preferences, contention between projects with different positive impacts will occur. To the extent that effort must be directed towards grant-seeking instead of scientific work, the resources available for the projects themselves are reduced, sometimes enough to cause a net loss.*

Remark 6. *Coordination limiting overuse of public goods is a major area of research in economics. Ostrom explains how such coordination is only possible when conflicts are anticipated or noticed and where a reliable mechanism can be devised [42].*

Model. *2.2—**Unnecessary Resource Contention.** As above, but each agent has an identical reward function of $f_{i,n}$. Even though all goals are shared, a lack of coordination in the above case leads to overspending, as shown in simple systems and for specified algebraic objective functions in the context of welfare economics. This literature shows many methods for how gains are possible, and in the simplest examples this occurs when agents coordinate to minimize overall spending on resource acquisition.*

Remark 7. *Coordination mechanisms themselves can be exploited by agents. The field of algorithmic game theory has several results for why this is only sometimes possible, and how building mechanisms to avoid such exploitation is possible [43].*

Failure Mode 3. ***Adversarial optimization** can occur when a victim agent has an incomplete model of how an opponent can influence the system. The opponent's model of the victim allows it to intentionally select for cases where the victim's model performs poorly and/or promotes the opponent's goal [3].*

Model. *3.1—**Adversarial Goal Poisoning.***

$$\begin{aligned} G_V &= x & M_V &= X : X \sim normal(x, \sigma^2(y)) \\ G_O &= -x & M_O &= (X, y) \end{aligned} \tag{4}$$

In this case, the Opponent O can see the metric for the victim, and can select for cases where y is large and X is small, so that V chooses maximal values of X, to the marginal benefit of O.

Example 4. *A victim's model can be learned by "Stealing" models using techniques such as those explored by Tramèr et al. [44]. In such a case, the information gained can be used for model evasion and other attacks mentioned there.*

Example 5. *Chess and other game engines may adaptively learn and choose openings or strategies for which the victim is weakest.*

Example 6. *Sophisticated financial actors can make trades to dupe victims into buying or selling an asset ("Momentum Ignition") in order to exploit the resulting price changes [45], leading to a failure of the exploited agent due to an actual change in the system which it misinterprets.*

Remark 8. *The probability of exploitable reward functions increases with the complexity of the system the agents manipulate [5], and the simplicity of the agent and their reward function. The potential for exploitation by other agents seems to follow the same pattern, where simple agents will be manipulated by agents with more accurate opponent models.*

Model. ***3.2—Adversarial Optimization Theft.*** *An attacker can discover exploitable quirks in the goal function to make the victim agent optimize for a new goal, as in Manheim and Garrabrant's Campbell's law example, slightly adapted here [3].*

$$\begin{aligned} M_V &= G_V + X \\ M_O &= G_O \cdot X \end{aligned} \tag{5}$$

O selects M_O after seeing V's choice of metric. In this case, we can assume the opponent chooses a metric to maximize based on the system and the victim's goal, which is known to the attacker. The opponent can choose their M_O so that the victim's later selection then induces a relationship between X and the opponent goal, especially at the extremes. Here, the opponent selects such that even weak selection on M_O hijacks the victim's selection on M_V to achieve their goal, because states where M_V is high have changed. In the example given, if $X \sim normal(\mu, \sigma^2)$, the correlation between G_O and M_O is zero over the full set of states, but becomes positive on the subspace selected by the victim. (Please note that the opponent choice of metric is not itself a useful proxy for their goal absent the victim's actions—it is a purely parasitic choice.)

Failure Mode 4. ***Input spoofing and filtering****—Filtered evidence can be provided, or false evidence can be manufactured and put into the training data stream of a victim agent.*

Model. ***4.1—Input Spoofing.*** *Victim agent receives public data $D(x_i|t)$ about the present world-state, and builds a model to choose actions which return rewards $f(x|t)$. The opponent can generate events x_i to poison the victim's learned model.*

Remark 9. *See the classes of data poisoning attacks explored by Wang and Chaudhuri [46] against online learning, and of Chen et al [47]. for creating backdoors in deep-learning verification systems.*

Example 7. *Financial market participants can (illegally) spoof by posting orders that will quickly be canceled in a "momentum ignition" strategy to lure others into buying or selling, as has been alleged to be occurring in high-frequency-trading [45]. This differs from the earlier example in that the transactions are not bona-fide transactions which fool other agents, but are actually false evidence.*

Example 8. *Rating systems can be attacked by inputting false reviews into a system, or by discouraging reviews by those likely to be the least or most satisfied reviewers.*

Model. ***4.2—Active Input Spoofing.*** *As in (4.1), where the victim agent employs active learning. In this case, the opponent can potentially fool the system into collecting data that seems very useful to the victim from crafted poisoned sources.*

Example 9. *Honeypots can be placed, or Sybil attacks mounted by opponents to fool victims into learning from examples that systematically differ from the true distribution.*

Example 10. *Comments by users "Max" and "Vincent DeBacco" on Eisen's blog post about Amazon pricing suggested that it is very possible to abuse badly built linear pricing models on Amazon to receive discounts, if the algorithms choose prices based on other quoted prices [38].*

Model. ***4.3—Input Filtering.*** *As in (4.1), but instead of generating false evidence, true evidence is hidden to systematically alter the distribution of events seen.*

Example 11. *Financial actors can filter the evidence available to other agents by performing transactions they do not want seen as private transactions or dark pool transactions.*

Remark 10. *There are classes of system where it is impossible to generate arbitrary false data points, but selective filtering can have similar effects.*

Failure Mode 5. ***Goal co-option*** *is when an opponent controls the system the Victim runs on, or relies on, and can therefore make changes to affect the victim's actions.*

Remark 11. *Whenever the computer systems running AI and ML systems are themselves insecure, it presents a very tempting weak point that potentially requires much less effort than earlier methods of fooling the system.*

Model. ***5.1—External Reward Function Modification.*** *Opponent O directly modifies Victim V's reward function to achieve a different objective than the one originally specified.*

Remark 12. *Slight changes in a reward function may have non-obvious impacts until after the system is deployed.*

Model. ***5.2—Output Interception.*** *Opponent O intercepts and modifies Victim V's output.*

Model. ***5.3—Data or Label Interception.*** *Opponent O modifies externally stored scoring rules (labels) or data inputs provided to Victim V's output.*

Example 12. *Xiao, Xiao, and Eckert explore a "label flipping" attack against support vector machines [48] where modifying a limited number of labels used in the training set can cause performance to deteriorate severely.*

Remark 13. *As noted above, there are cases where generating false data may be impossible or easily detected. Modifying the inputs during training may create less obvious traces of an attack has occurred. Where this is impossible, access can also allow pure observation which, while not itself an attack, can allow an opponent to engage in various other exploits discussed earlier.*

To conclude the list of failure modes, it is useful to note a few areas where the failures are induced or amplified. This is when agents explicitly incentivize certain behaviors on the part of other agents, perhaps by providing payments. These public interactions and incentive payments are not fundamentally different from other failure modes, but can create or magnify any of the other modes. This is discussed in literature on the evolution of collusion, such as Dixon's treatment [49]. Contra Dixon, however, the failure modes discussed here can prevent the collusion from being beneficial. A second, related case is when creating incentives where an agent fails to anticipate either the ways in which the other agents can achieve the incentivized target, or the systemic changes that are induced. These so-called "Cobra effects" [3] can lead to both the simpler failures of the single-agent cases explored in Manheim and Garrabrant, and lead to the failures above. Lastly, as noted by Sandberg [50], agents with different "speeds" (and, equivalently, processing power per unit time,) can exacerbate victimization, since older and slower systems are susceptible, and susceptibility to attacks only grows as new methods of exploitation are found.

3. Discussion

Multi-agent systems can naturally give rise to cooperation instead of competition, as discussed in Leibo et al.'s 2017 paper [51]. The conditions under which there is exploitation rather than cooperation, however, are less well understood. A more recent paper by Leibo proposes that the competition dynamic can be used to encourage more complex models. This discusses coordination failures, but the discussion of dynamics leading to the failures does not engage with the literature on safety or goal-alignment [52]. Leibo's work, however, differs from most earlier work where multi-agent systems are trained together with a single goal, perforce leading to cooperative behavior, as in Lowe et al.'s heavily cited work, in which "competitive" dynamics are dealt with by pre-programming explicit models of other agent behaviors [53].

The failure modes outlined (accidental steering, coordination failures, adversarial misalignment, input spoofing or filtering, and goal co-option or direct hacking) are all due to models that do not fully account for other agent behavior. Because all models must simplify the systems they represent, the prerequisites for these failures are necessarily present in complex-enough systems where multiple non-coordinated agents interact. The problems of embedded agents discussed by Demski and Garrabrant [36] make it particularly clear that current approaches are fundamentally unable to fully represent these factors. For this and other reasons, mitigating the failures modes discussed here are not yet central to the work of building better ML or narrow AI systems. At the same time, some competitive domains such as finance are already experiencing some of these exploitative failures [45], and bots engaging in social network manipulation, or various forms of more direct interstate competition are likely engaging in similar strategies.

The failures seen so far are minimally disruptive. At the same time, many of the outlined failures are more problematic for agents with a higher degree of sophistication, so they should be expected not to lead to catastrophic failures given the types of fairly rudimentary agents currently being deployed. For this reason, specification gaming currently appears to be a mitigable problem, or as Stuart Russell claimed, be thought of as "errors in specifying the objective, period" [54]. This might be taken to imply that these failures are avoidable, but the current trajectory of these systems means that the problems will inevitably worsen as they become more complex and more such systems are deployed, and the approaches used are fundamentally incapable of overcoming the obstacles discussed.

Potential Avenues for Mitigation

Mitigations for these failures exist, but as long as the fundamental problems discussed by Demski and Garrabrant [36] are unaddressed, the dynamics driving these classes of failure seem unavoidable. Furthermore, such failures are likely to be surprising. They will emerge as multiple machine learning agents are deployed, and more sophisticated models will be more likely to trigger them. However, as argued above, these failures are fundamental to interaction between complex agents. This means that while it is unclear how quickly such failures will emerge, or if they will be quickly recognized, it is unquestionable that they will continue to occur. System designers and policymakers should expect that these problems will become intractable if deferred, and are therefore particularly critical to address now. It is be expected that any solution involves a combination of approaches [17], though the brief overview of safety approaches below shows that not all general approaches to AI safety are helpful for multi-agent failures.

First, there are approaches that limit optimization. This can be done via satisficing, using approaches such as Taylor's Quantilizers, which pick actions at random from the top quantile of evaluated choices [55]. Satisficing approaches can help in prevent exploiting other agents, or in preventing accidental overoptimization, but are not effective as a defense against exploitative agents or systemic failures due to agent interaction. Another approach limiting optimization is explicit safety guarantees. In extrema, this looks like an AI-Box, preventing any interaction of the AI with the wider world and hence preventing agent interaction completely. This is effective if such boxes are not escaped, but it is unclear if this is possible [27]. Less extreme versions of safety guarantees are sometimes possible, especially in domains where a formal model of safe behavior is possible, and the system is sufficiently well understood. For example, Shalev-Shwartz et al. have such a model for self-driving cars, heavily relying on the fact that the physics involved with keeping cars from hitting one another, or other objects, is in effect perfectly understood [56]. Expanding this to less well understood domains seems possible, but is problematic for reasons discussed elsewhere [57].

Without limiting optimization explicitly, some approaches attempt to better define the goals, and thereby reduce the extent of unanticipated behaviors. These approaches involve some version of direct optimization safety. One promising direction for limiting the extent to which goal-directed optimization can be misdirected is to try to recognize actions rather than goals [58]. Human-in-the-loop oversight is another direction for minimizing surprise and ensuring alignment, though this is already infeasible in many systems [20]. Neither approach is likely to be more effective than humans themselves are at preventing such exploitation. The primary forward-looking approach for safety is some version of ensuring that the goal is aligned, which is the bulk of what Yampolskiy and Fox refer to as AI safety engineering [59].

In multi-agent contexts there is still a concern that because human values are complex, [18] exploitation is an intrinsically unavoidable pitfall in multi-agent systems. Paul Christiano's "Distillation and Amplification" approach involves safe amplification using coordinated multi-agent systems [60]. This itself involves addressing some of the challenges with multi-agent approaches, and work on safe amplification using coordinated multi-agent systems in that context has begun [61]. In that work, the coordinating agents are predictive instead of agentic, so the failure modes are more restricted. The methods suggested can also be extended to agentic systems, where they may prove more worrisome, and solving the challenges potentially involves mitigating several failure modes outlined here.

Between optimization-limiting approaches and AI safety engineering, it is possible that many of the multi-agent failures discussed in the paper can be mitigated, though not eliminated. In addition, there will always be pressure to prioritize performance as opposed to safety, and safe systems are unlikely to perform as quickly as unsafe ones [20]. Even if the tradeoff resolves in favor of slower, safer systems, such systems can only be created if these approaches are further explored and the many challenges involved are solved before widespread deployment of unsafe ML and AI. Once the systems are deployed, it seems infeasible that safer approaches could stop failures due to exploiting and exploitable systems, short of recalling them. This is not a concern for the far-off future where misaligned superintelligent AI poses an existential risk. It is instead a present problem, and it is growing more serious along with the growth of research that does not address it.

4. Conclusions: Model Failures and Policy Failures

Work addressing the failure modes outlined in the paper is potentially very valuable, in part because these failure modes are mitigable or avoidable if anticipated. AI and ML system designers and users should expect that many currently successful but naive agents will be exploited in the future. Because of this, the failure modes are likely to become more difficult to address if deferred, and are therefore particularly critical to understand and address them preemptively. This may take the form of systemic changes such as redesigned financial market structures, or may involve ensuring that agents have built-in failsafes, or that they fail gracefully when exploited.

At present, it seems unlikely that large enough and detected failures will be sufficient to slow the deployment of these systems. It is possible that governmental actors, policymakers, and commercial entities will recognize the tremendous complexities of multiparty coordination among autonomous agents and address these failure modes, or slow deployment and work towards addressing these problems even before they become catastrophic. Alternatively, it is possible these challenges will become apparent via limited catastrophes that are so blatant that AI safety will be prioritized. This depends on how critical the failures are, how clearly they can be diagnosed, and whether the public demands they be addressed.

Even if AI amplification remains wholly infeasible, humanity is already deploying autonomous systems with little regards to safety. The depth of complexity is significant but limited in current systems, and the strategic interactions of autonomous systems are therefore even more limited. However, just as AI for poker eventually became capable enough to understand multi-player interaction and engage in strategic play, AI in other systems should expect to be confronted with these challenges. We do not know when the card sharks will show up, or the extent to which they will make the games they play unsafe for others, but we should admit now that we are as-yet unprepared for them.

Funding: This research was funded in large part by a grant from the Berkeley Existential Risk Initiative.

Acknowledgments: I would like to thank a subset of the anonymous reviewers in both the first and second submission for very helpful comments, and thank Roman Yampolskiy for encouraging me to write and revise the paper, despite setbacks.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the writing of the manuscript, nor in the decision to publish the results.

References

1. Clark, J.; Amodei, D. Faulty Reward Functions in the Wild. 2016. Available online: https://openai.com/blog/faulty-reward-functions/ (accessed on 12 March 2019).
2. Goodhart, C.A.E. *Problems of Monetary Management: The UK Experience*; Papers in Monetary Economics; Reserve Bank of Australia: Sydney, Australia, 1975.
3. Manheim, D.; Garrabrant, S. Categorizing Variants of Goodhart's Law. *arXiv* **2018**, arXiv:1803.04585.
4. Campbell, D.T. Assessing the impact of planned social change. *Eval. Program Plan.* **1979**, *2*, 67–90. [CrossRef]
5. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete problems in AI safety. *arXiv* **2016**, arXiv:1606.06565.
6. Kleinberg, J.; Raghavan, M. How Do Classifiers Induce Agents To Invest Effort Strategically? *arXiv* **2018**, arXiv:1807.05307.
7. Bostrom, N. *Superintelligence*; Oxford University Press: Oxford, UK, 2017.
8. Braganza, O. Proxyeconomics, An agent based model of Campbell's law in competitive societal systems. *arXiv* **2018**, arXiv:1803.00345.
9. Krakovna, V. Specification Gaming Examples in AI. 2018. Available online: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/ (accessed on 12 March 2019).
10. Liu, L.; Cheng, L.; Liu, Y.; Jia, Y.; Rosenblum, D.S. Recognizing Complex Activities by a Probabilistic Interval-based Model. In Proceedings of the National Conference on Artificial Intelligence (AAAI), Phoenix, AZ, USA, 12–17 February 2016.
11. Cheney, N.; MacCurdy, R.; Clune, J.; Lipson, H. Unshackling evolution: Evolving soft robots with multiple materials and a powerful generative encoding. *ACM SIGEVOlution* **2014**, *7*, 11–23. [CrossRef]
12. Figueras, J. Genetic Algorithm Physics Exploiting. 2015. Available online: https://youtu.be/ppf3VqpsryU (accessed on 12 Mach 2019).
13. Lehman, J.; Clune, J.; Misevic, D.; Adami, C.; Beaulieu, J.; Bentley, P.J.; Bernard, S.; Belson, G.; Bryson, D.M.; Cheney, N. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *arXiv* **2018**, arXiv:1803.03453.
14. Chopra, J. GitHub issue for OpenAI gym environment FetchPush-v0. 2018. Available online: https://github.com/openai/gym/issues/920 (accessed on 12 Mach 2019).
15. Popov, I.; Heess, N.; Lillicrap, T.; Hafner, R.; Barth-Maron, G.; Vecerik, M.; Lampe, T.; Tassa, Y.; Erez, T.; Riedmiller, M. Data-efficient deep reinforcement learning for dexterous manipulation. *arXiv* **2017**, arXiv:1704.03073.
16. Weimer, W. Advances in Automated Program Repair and a Call to Arms. In *Proceedings of the 5th International Symposium on Search Based Software Engineering—Volume 8084*; Springer: Berlin/Heidelberg, Germany, 2013.
17. Sandberg, A. Friendly Superintelligence. Presentation at Extro 5 Conference. Available online: http://www.nada.kth.se/~asa/Extro5/Friendly%20Superintelligence.htm, 2001. (accessed on 12 March 2019).
18. Yudkowsky, E. Complex value systems in friendly AI. In Proceedings of the International Conference on Artificial General Intelligence, Mountain View, CA, USA, 3–6 August 2011; Springer: New York, NY, USA, 2011; pp. 388–393.
19. Worley, G.G., III. Robustness to fundamental uncertainty in AGI alignment. *arXiv* **2018**, arXiv:1807.09836.
20. Danzig, R. *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*; Technical Report; Center for a New American Security: Washington, DC, USA, 2018.
21. Baum, S. Superintelligence skepticism as a political tool. *Information* **2018**, *9*, 209. [CrossRef]
22. Yudkowsky, E. Intelligence explosion microeconomics. *Mach. Intell. Res.* **2013**, *23*, 2015.
23. Lewis, G.T. Why the Tails Come Apart Apart. Lesswrong, 2014. Available online: http://lesswrong.com/lw/km6/whythetailscomeapart/ (accessed on 12 March 2019).
24. Yudkowsky, E. *The AI Alignment Problem: Why It's Hard, and Where to Start*; Stanford University: Stanford, CA, USA, 2016.
25. Behzadan, V.; Munir, A. Models and Framework for Adversarial Attacks on Complex Adaptive Systems. *arXiv* **2017**, arXiv:1709.04137.

26. Drexler, K.E. *Engines of Creation*; Anchor: New York, NY, USA, 1986.
27. Armstrong, S.; Sandberg, A.; Bostrom, N. Thinking inside the box: Controlling and using an oracle AI. *Minds Mach.* **2012**, *22*, 299–324. [CrossRef]
28. Mulligan, T.S. How Enron Manipulated State's Power Market. *Los Angeles Times*, 9 May 2002. Available online: http://articles.latimes.com/2002/may/09/business/fi-scheme9 (accessed on 9 March 2019).
29. Borel, E.; Ville, J. *Applications de la théorie des Probabilités aux jeux de Hasard*; Gauthier-Villars: Paris, France, 1938.
30. Kuhn, H.W. A simplified two-person poker. *Contrib. Theory Games* **1950**, *1*, 97–103.
31. Bowling, M.; Burch, N.; Johanson, M.; Tammelin, O. Heads-up limit hold'em poker is solved. *Science* **2015**, *347*, 145–149. [CrossRef] [PubMed]
32. Brown, N.; Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science* **2018**, *359*, 418–424. [CrossRef] [PubMed]
33. Billings, D.; Burch, N.; Davidson, A.; Holte, R.; Schaeffer, J.; Schauenberg, T.; Szafron, D. Approximating game-theoretic optimal strategies for full-scale poker. *IJCAI* **2003**, *3*, 661.
34. Soares, N. Formalizing Two Problems of Realistic World-Models. Technical Report. Available online: https://intelligence.org/files/RealisticWorldModels.pdf (accessed on 9 March 2019).
35. Conant, R.C.; Ross Ashby, W. Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* **1970**, *1*, 89–97. [CrossRef]
36. Demski, A.; Garrabrant, S. Embedded Agency. *arXiv* **2019**, arXiv:1902.09469.
37. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Broadway Books: New York City, NY, USA, 2016.
38. Eisen, M. Amazon's $23,698,655.93 Book about Flies. 2011. Available online: http://www.michaeleisen.org/blog/?p=358 (accessed on 9 March 2019).
39. Smaldino, P.E.; McElreath, R. The natural selection of bad science. *Open Sci.* **2016**, *3*, 160384. [CrossRef] [PubMed]
40. Gibbard, A. Manipulation of Voting Schemes: A General Result. *Econometrica* **1973**, *41*, 587–601. [CrossRef]
41. Yudkowsky, E. *Inadequate Equilibria: Where and How Civilizations Get Stuck*; Machine Intelligence Research Institute: Berkeley, CA, USA, 2017.
42. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action*; Cambridge University Press: Cambridge, UK, 1990.
43. Nisan, N.; Roughgarden, T.; Tardos, E.; Vazirani, V.V. *Algorithmic Game Theory*; Cambridge University Press: Cambridge, UK, 2007.
44. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing Machine Learning Models via Prediction APIs. In Proceedings of the USENIX Security Symposium, Vancouver, BC, Canada, 16–18 August 2016; pp. 601–618.
45. Shorter, G.W.; Miller, R.S. *High-Frequency Trading: Background, Concerns, and Regulatory Developments*; Congressional Research Service: Washington, DC, USA, 2014; Volume 29.
46. Wang, Y.; Chaudhuri, K. Data Poisoning Attacks against Online Learning. *arXiv* **2018**, arXiv:1808.08994.
47. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv* **2017**, arXiv:1712.05526.
48. Xiao, H.; Xiao, H.; Eckert, C. Adversarial Label Flips Attack on Support Vector Machines. *Front. Artif. Intell. Appl.* **2012**, *242*, doi:10.3233/978-1-61499-098-7-870. [CrossRef]
49. Dixon, H.D. Keeping up with the Joneses: Competition and the evolution of collusion. *J. Econ. Behav. Organ.* **2000**, *43*, 223–238. [CrossRef]
50. Sandberg, A. There is plenty of time at the bottom: The economics, risk and ethics of time compression. *Foresight* **2018**, *21*, 84–99. [CrossRef]
51. Leibo, J.Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; Graepel, T. Multi-agent reinforcement learning in sequential social dilemmas. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, São Paulo, Brazil, 8–12 May 2017; pp. 464–473.
52. Leibo, J.Z.; Hughes, E.; Lanctot, M.; Graepel, T. Autocurricula and the Emergence of Innovation from Social Interaction: A Manifesto for Multi-Agent Intelligence Research. *arXiv* **2019**, arXiv:1903.00742.
53. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, O.P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. *Adv. Neural Inf. Process. Syst.* **2017**, 6379–6390.

54. Russell, S. Comment to Victoria Krakovna, Specification Gaming Examples in AI. 2018. Available online: https://perma.cc/3U33-W8HN (accessed on 12 March 2019).
55. Taylor, J. Quantilizers: A safer alternative to maximizers for limited optimization. In Proceedings of the Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
56. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. On a formal model of safe and scalable self-driving cars. *arXiv* **2017**, arXiv:1708.06374.
57. Manheim, D. Oversight of Unsafe Systems via Dynamic Safety Envelopes. *arXiv* **2018**, arXiv:1811.09246.
58. Liu, Y.; Nie, L.; Liu, L.; Rosenblum, D.S. From action to activity: Sensor-based activity recognition. *Neurocomputing* **2016**, *181*, 108–115. [CrossRef]
59. Yampolskiy, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
60. Christiano, P.; Shlegeris, B.; Amodei, D. Supervising strong learners by amplifying weak experts. *arXiv* **2018**, arXiv:1810.08575.
61. Irving, G.; Christiano, P.; Amodei, D. AI safety via debate. *arXiv* **2018**, arXiv:1805.00899.

big data and cognitive computing

MDPI

Article

Global Solutions vs. Local Solutions for the AI Safety Problem

Alexey Turchin [1,*], David Denkenberger [2] and Brian Patrick Green [3]

[1] Science for Life Extension Foundation, Prospect Mira 124-15, Moscow 129164, Russia
[2] Alliance to Feed the Earth in Disasters (ALLFED), University of Alaska Fairbanks, Fairbanks, AK 99775, USA; ddenkenberger@alaska.edu
[3] Markkula Center for Applied Ethics, Santa Clara University, Santa Clara, CA 95053, USA; bpgreen@scu.edu
[*] Correspondence: alexeiturchin@gmail.com

Received: 16 December 2018; Accepted: 15 February 2019; Published: 20 February 2019

Abstract: There are two types of artificial general intelligence (AGI) safety solutions: global and local. Most previously suggested solutions are local: they explain how to align or "box" a specific AI (Artificial Intelligence), but do not explain how to prevent the creation of dangerous AI in other places. Global solutions are those that ensure any AI on Earth is not dangerous. The number of suggested global solutions is much smaller than the number of proposed local solutions. Global solutions can be divided into four groups: 1. No AI: AGI technology is banned or its use is otherwise prevented; 2. One AI: the first superintelligent AI is used to prevent the creation of any others; 3. Net of AIs as AI police: a balance is created between many AIs, so they evolve as a net and can prevent any rogue AI from taking over the world; 4. Humans inside AI: humans are augmented or part of AI. We explore many ideas, both old and new, regarding global solutions for AI safety. They include changing the number of AI teams, different forms of "AI Nanny" (non-self-improving global control AI system able to prevent creation of dangerous AIs), selling AI safety solutions, and sending messages to future AI. Not every local solution scales to a global solution or does it ethically and safely. The choice of the best local solution should include understanding of the ways in which it will be scaled up. Human-AI teams or a superintelligent AI Service as suggested by Drexler may be examples of such ethically scalable local solutions, but the final choice depends on some unknown variables such as the speed of AI progress.

Keywords: AI safety; existential risk; AI alignment; superintelligence; AI arms race

1. Introduction

The problem of how to prevent a global catastrophe associated with the expected development of AI of above human-level intelligence is often characterized as "AI safety" [1]. The topic has been explored by many researchers [2–5]. Other forms of "AI safety," typically associated with narrow-AI such as that for self-driving cars or other narrow applications, are not considered in this paper.

An extensive review of possible AI safety solutions has been conducted by Sotala and Yampolskiy [4]. In their article, they explore a classification of AI safety solutions by social, external, and internal measures.

In this article, we suggest a different classification of AI safety solutions, as *local* or *global*, and describe only global solutions. Local solutions are those that affect only *one* AI, and include AI ethics, AI alignment, AI boxing, etc. Global solutions are those that affect any potential AI in the world, for example, global technological relinquishment or use of the first superintelligent AI to prevent other AIs from arising. Most solutions described by Sotala and Yampolskiy [4] are considered local solutions in our classification scheme.

Recent significant contributions to the global solutions problem include Christiano's slow takeoff model [6], which demonstrated that such a takeoff could happen earlier than a fast takeoff; Ramamoorthy and Yampolskiy's research on AI arms races: "Beyond MAD?: the race for artificial general intelligence" [7]; Brundage et al.'s "The Malicious Use of Artificial Intelligence" [8]; and research on a collective takeoff by Sotala [9]. The problem of "other AIs," central to the global AI safety conundrum, has been explored by Dewey [10], who suggested four types of solution: international coordination, sovereign AI (superintelligent AI acting independently on global scale), an AI-empowered project, and some other decisive technological advantage.

Any local safety solution which cannot be applied globally cannot in itself determine the course of human history, as many other AIs may appear with different local properties. However, some local solutions could reach the global level if an external transfer mechanism is added, such as an international agreement, or if the first AI based on this local solution becomes the only global power: *Singleton* [11].

Generally, when we use the term "AI" throughout this article, we do not mean standard contemporary systems of machine learning processing Big Data, but rather the descendants of these contemporary systems, which are dramatically more sophisticated, nuanced, and process vastly more data even faster, and therefore attain intelligence equivalent to and/or surpassing human intelligence. Additionally, this article is based on the assumption—shared by many (e.g., [3,12]), but not all AI researchers—that above human-level AI is possible in the relatively near future (21st century) and the world's socio-political structure will be approximately the same as now at the moment of its creation. This assumption about the possibility of superhuman AI is naturally followed by concerns about the safety of such systems, which may generate not only isolated accidents, but a full variety of possible global catastrophes as explored in Reference [13].

The main thesis of this article is that there are two main types of AI safety solutions: local and global, and that not every local solution scales to a global solution or does it ethically and safely. The choice of the best local solution should include an understanding of the ways in which it may be scaled up. Human-AI teams or a superintelligent AI Service as suggested by Drexler [14] may be examples of such ethically scalable local solutions, but the final choice depends on some unknown variables such as the speed of AI progress [15].

To solve the problem of the relation between global and local solutions, we created a classification of global solutions, which is a simpler task as all global solutions depend on the one main variable: how many different AI systems will be eventually created. We used this classification to identify pairs of local and global solutions, which are less risky when combined.

In Section 2 we overview various levels of AI safety. In Section 3 we look at solutions involving the prevention of AI, while in Section 4 we explore "one AI solutions," where the first AI prevents the appearance of other AIs. In Section 5 we address "many AI solutions," in which many superhuman AIs appear and interact. In Section 6, we suggest a class of solutions in which technologically modified human beings or human-mind models collaborate directly with or control AI, "inside" it.

2. AI Safety Levels

To explore how to implement a global AI safety solution, we need some insight about what human safety may look like in the future. Global human safety in the far future [16] may be reached at different levels, from miserable survival to extreme flourishing. According Bostrom's classification, everything below full realization of the human potential is an existential risk [17], but low realization is not the same as extinction [18], and Green argues that even full human flourishing is not enough to eliminate existential risk unless ethical standards and practices are also somehow concomitantly perfected [19].

Several preliminary levels of AI safety may be suggested, similar to the classification of AI safety levels presented in a report from the Foundational Research Institute by Brian Tomasik [20], but centered on suffering. Our classification is based on levels of human well-being, the first and most basic of which is survival:

1. "Last man": At least one human being survives creation of strong AI, for example, possibly as an upload, or in a "zoo".
2. "AI survivors": A group of people survive and continue to exist after the AI creation, and may be able to rebuild human civilization to some extent. This may happen if the AI halts [13] or leaves Earth.
3. "No AI": Any outcome where global catastrophe connected with AI has not occurred because there is no AI to provoke the catastrophe. This is a world in which a comprehensive ban on AI is enforced, or AI technologies otherwise never progress to AGI or superintelligence.
4. "Better now": Human civilization is preserved after AI creation in almost the same form in which it exists now, and benefits from AI in many ways, including through the curing of diseases, slowing aging, preventing crime, increasing material goods, achieving interstellar travel, etc. Outcomes in this category likely involve a type of "AI Nanny" [21].
5. "Infinite good": Superintelligent AI which maximizes human values (Benevolent AI) will reach the maximum possible positive utility for humans, but contemporary humans cannot now describe this utility as it is beyond our ability to imagine, as presented by Yudkowsky [22].

Different global solutions of the AI safety problem provide different levels of survival as the most plausible outcome. From our point of view, Levels 3, 4 and 5 are acceptable outcomes, and Levels 1 and 2 are unacceptable as they produce unimaginable human suffering and risk human extinction.

3. "No AI" Solutions

In our world of quick AI development, AI relinquishment seems improbable or requires some unethical and/or risky acts of Luddism. Many of these solutions have been explored by Sotala amd Yampolskiy [4].

Overview of restrictive solutions where advance AI creation is prevented globally:

- International ban
- Legal relinquishment
- Technical relinquishment or AI appears to be not technically possible
- Destruction of capability to produce AI anywhere in the world
  - War
  - Luddism
  - Staging small catastrophe
- Slowdown of AI creation
  - Economical
  - Technology slowdown
  - Overregulation
  - Brain drain from the field
  - Defamation of idea of AI, AI winter

3.1. Legal Solutions, Including Bans

Not many argue for a global AI ban as it is unfeasible under current conditions [23] and would likely only help bad actors [24]. One could imagine that global legal regulation could ban the creation of self-improving agents. However, in our current, divided world its enforcement would be difficult. Only a powerful global government could make such a solution workable.

Some form of regulation may appear ad hoc, as an urgent measure implemented by the UN, or a group of the most powerful countries. However, they would need very credible harbingers as motivation. These could be several epidemics of AI-viruses of increasing strength, i.e., computer viruses with elements of machine learning [13]. However, there is currently no agreement as to what factors would serve as a credible "alarm," and such agreement may be impossible [25].

Some governmental and non-governmental groups are working to develop guidelines in this area. The EU is considering legislation about robotic ethics [26]. Similar legislation may ban potentially dangerous self-improving systems, and if adopted in the most developed countries, it may act as a proxy for a global ban. It could be enforced in smaller, rogue countries by military coalitions, similar to the one formed in the 2003 Iraq war, but such a ban cannot be created and enforced without understanding the risks of AI. The recent Asilomar AI Guidelines [27] could also serve as a foundation for internal control within the AI community to prevent creation of recursively self-improving (RSI) AI. The Asilomar guidelines could also form the basis for international law regulating AI.

Elon Musk recently advocated global regulation of AI research [28]. Such regulations may take the form of a UN agency similar to the International Atomic Energy Agency (IAEA). The IAEA provides safety protocols for its members, demands openness and conducts inspections to confirm implementation; in exchange, it gives access to recent results on other members. The result will be something similar to Open AI, as described by Reference [29], but enforced by the UN.

To implement such an AI agency, the UN would need a powerful enforcement agency. In the same way as when the IAEA fails, an international coalition would need to be able to use sanctions (as against Iran and North Korea) or military intervention, as in Iraq. However, a UN-backed AI-control agency would require much tighter and swifter control mechanisms, and would be functionally equivalent to a world government designed specifically to contain AI. To be effective, such an agency must be empowered to use force, possibly including cyber weapons or possibly even nuclear weapons. However, in the current world climate, there will be little or no support for the creation of a world government authorized to use powerful weapons to destroy AI labs based only on theory. The only chance for its creation might be if some spectacular AI accident happened, for example, if a narrow-AI-based virus with machine learning capabilities hacked hundreds of airplanes and crashed them into nuclear power plants. In such a case, a global ban on advanced AI might be possible.

3.2. Restriction Solutions

The idea of restriction is to find a scarce "commodity" needed for the creation of AI and try to limit access to it [30–32]. A global authority would be needed to implement such bans.

Such "commodities" could include:

- supercomputers
- programmers
- knowledge about AI creation
- semiconductor fabrication plants ("fabs")
- internet access
- electricity

The rarest commodity are chip fabs, which cost billions of dollars and are needed to create new processors. There are around 200 chip fabs in the world now [33]. If they were closed, no new computers could appear in the world, which might drastically slow AI progress. However, the effect of fabs is rather indirect, as it is possible that enough computers already exist to create AI, especially given the large existing supply of graphics cards, but these chips are simply not in the right configuration.

Large datacenters, supercomputers, scientific centers, and internet hubs are also relatively rare, with the number worldwide in the thousands. Current home PCs (not connected to a network) are probably unable to support AI, so if powerful computers and internet connections are switched off, it could considerably slow down AI creation. These restrictions, such as those discussed in Section 3.1, would require preexisting global coordination. In addition, they would obviously have significant economic consequences.

As a last point, in this section it is worth noting that full AI may not be technically possible in any realistic sense, and if that is the case then whatever "commodity" permits its creation is restricted in a complete sense.

3.3. Destructive Solutions

One possible way to stop the creation of AI is annihilation by a nuclear attack of AI research centers, electronic equipment, and sources of electricity, which could be done locally or globally by a nuclear country acting alone (a conventional attack could also be attempted, but would be slower and have a lower probability of success). If such an attack was carried out against an adversary, it would "just" be a war; if done globally, it would mean that a superpower would bomb its own AI labs. Nuclear attack of this type is extremely unlikely, unless it were perceived that an "AI uprising" had already started.

Similar to the first option, but with a more purely anti-electronics approach, destruction could be accomplished by a multitude of high-altitude electromagnetic pulses (HEMPs) caused by nuclear detonations. A concerted attack of this kind could destroy all unshielded electronics. Because electricity, fossil fuel extraction, and industry, all depend on electronics, manufacturing and distribution would grind to a halt. This would not kill people directly, but could cause mass human starvation unless society were prepared [34,35]. However, recovery of technological civilization and thus the ability to recreate AI is possible, so the problem would probably appear again. Alternatively, chaos could result in a downward spiral leading to extinction. So, it is a risky "solution" that, even if it succeeded, would likely be temporary.

One could imagine other means of destruction, ranging from economic recession to Luddism [36], to various global catastrophes, but, as with the above options, all of them are impractical and morally unacceptable. In the future, perhaps some high-tech methods of AI halting might be implemented, such as the Stuxnet computer virus that destroyed Iran's uranium centrifuges [37]. A virus could be used to destroy chip fabs, shut down the internet, or cut electricity. There are other ideas in the field, but an exhaustive list is not within the scope of this paper.

As one last point, the unilateralist's curse—the lack of coordination between many actors with the same goal [38]—may exaggerate activities of those groups that at least believe in the possibility of safe AI.

3.4. Delay of AI Creation

The global recession of 2008 did not have any measurable effect on the speed of AI development. Only a large-scale economic collapse that significantly disrupted global trade could slow AI development to any significant extent.

Other events could slow down AI development, include:

- Public fears of AI.
- The next AI winter, lack of interest in its development (there have already been two after hype in the 1960s and 1980s).
- Extensive regulation of the field.
- Intentional disruption of the research field via fake news, defamation, white noise, and other instruments of informational warfare.
- Public ridicule of the field after some failure.
- Change of focus of public attention by substitution of terms. This happened with "nanotechnology", which originally meant a powerful manufacturing technology, but now means making anything small. Such a shift may happen with the term "AI," where the meaning has shifted recently from human-like systems to narrow machine learning algorithms. There are

several fields that have had slow development for decades because of marginalization, such as cryonics, but it looks as though the time of marginalization of AI has passed.

- Lastly, depending on the technical challenges, advanced AI, including AGI and superintelligence, may not be technically possible in the near future, although there is no reason at this point to assume it is not. However, if these challenges appear, AI could be indefinitely delayed.

4. ”One AI” Solutions

These solutions are centered on the idea that the first AI will become dominant and prevent the development of other AIs. The nature of these solutions is that they are implemented locally, but affect the whole globe due to the global power of the singleton.

Overview of “one AI” solutions:

- First AI is used to take over the world
  - First AI is used as a military instrument
  - First AI gains global power via peaceful means
    - Commercial success
    - Superhuman negotiating abilities
  - Strategic advantage achieved by narrow AIs produces global unification, before the rise of superintelligent AI, by leveraging preexisting advantage of a nuclear power and increasing first-strike capability
  - First AI is created by a superpower and provides it a decisive strategic advantage
  - First AI is reactive, and while it does not prevent the creation of other AI, it limits their potential danger
  - First AI is a genius at negotiation and solves all conflicts between other agents
- First AI appears as a result of collective efforts
  - AI police: global surveillance system to prevent creation of dangerous AI
  - “AI CERN”: international collaboration creates an AI Nanny
  - Main players collaborate with each other
  - AIs are effective in cooperation and merge with each other
- Non-agential AI-medium (AI as widely distributed technology, without agency)
  - Comprehensive AI Services
  - Distributed AI based on blockchain (SingularityNET)
  - AI as technology everywhere (openness)
  - Augmented humans as AI neurons (Neuralink)
  - Superintelligence as a distributed optimization process by rivalry between AI agents (market)

Indirect measures to increase probability that first AI will be human-aligned:

- Helping others to create safe first AI
  - AI safety theory is distributed among main players and used by every AI creator
  - AI safety instruments are sold as a service
  - Promotion of AI safety
- Slowing creation of other AIs
  - Concentrate best minds on other projects and remove them from AI research
  - Take low-hanging research fruit

- Factors affecting the arms race for AI include funding, openness, number of teams, prizes, and public attitudes

4.1. First AI Seizes World Power

Advanced agential AIs will be able to act in the world autonomously. Superintelligent AI could potentially seize world power on its own. Max Tegmark describes a scenario in which the first AI initially gains world dominance through earning money and later consolidates power by rigging elections or staging coups in different countries [39].

The main problem of the idea that first AI can be used as an instrument to take over the world is that it creates motivation for militarisation of AI, which has potentially dangerous consequences [40].

Superintelligent AI may be able to find win-win solutions in negotiations. Such an ability could help it overcome resistance to global unification, as it will be able to provide its unique negotiating ability as a service, which everyone will be interested in applying, and in that case, there will be no need for a military world takeover.

4.1.1. Concentrate the Best AI Researchers to Create a Powerful and Safe AI First

This idea is to create something similar to the Manhattan Project, attracting the best minds to work together on the creation of the first self-improving AI. This would provide such a large concentration of human intelligence that they could simultaneously create AI and solve the problem of AI safety. The Manhattan Project was formed of the best scientists in the world, and they were concerned about potential global risks of the first nuclear explosion. For example, scientists involved in the project created the LA-602 report about the possibility of causing a nuclear-initiated chain reaction in the atmosphere [41].

Later efforts to create nuclear weapons in other countries were not so safety-oriented. The Soviets exploded a bomb over their own troops [42]. The Indians dropped explosives intended to be part of their first nuclear bomb during critical assembly—fortunately, it did not detonate [43].

If a similar trend holds for AI research, the first concerted effort may be more safety-oriented and involve better planning and brighter minds than later efforts. In addition, if research is accelerated in one research institution, it could outperform the world in general. This could help prevent a troubling situation in which safety solutions are well-understood in one organization, but AI is created by another group.

If the first effort is ahead of the competitors by years, it will have a safety time gap, that is, additional time for working on AI safety. In other words, the leader would have more time to think about safety, by virtue of their being in the lead.

In early stages of its development (in the 2000s), the Machine Intelligence Research Institute (MIRI) had a plan to be the creator of the first Friendly AI. However, its goal now is to facilitate research on AI safety solutions [44,45] to be implemented elsewhere.

4.1.2. Using the Decisive Advantage of Non-Self-Improving AI to Create an AI Nanny

Sotala [46] wrote that even non-self-improving AI may gain a decisive strategic advantage if it is effective at designing new weapons, or in strategic military or political planning. This opens the possibility to use the first human-level AI to gain power over the world, without taking the dangerous and unpredictable route of recursive self-improvement.

Such AI might be built around a human upload or its equivalent, which gains most of its power not from self-improvement, but from running on high-speed hardware. Such a high-speed human analogue gaining global power via social manipulation and designing new weapons might become an "AI king".

One way to gain such a decisive strategic advantage would be if the first AI were created by a superpower (either China or the US) which is already close to world domination. Such an AI, created as a government-sponsored large project, may be attained as part of a secret "Manhattan Project"-type effort or by seizing the archives and work of a large private company. The AI could leverage other power-projecting instruments already controlled by this superpower to provide it with the capability for world domination (e.g., access to secret information, control of nuclear weapons, large financial resources). Exemplifying this view, see the recent remark by Vladimir Putin that "the nation that leads in AI 'will be the ruler of the world'" [47].

For example, even narrow-AI designed to calculate nuclear war scenarios could provide a decisive strategic advantage for an existing nuclear superpower. It could then strike in a way that yields a high probability of no retaliation.

Dewey [10] suggested the first AI could be reactive or proactive: Proactive AI prevents creation of other AIs, starting preemptive wars against them, and reactive AI only limits or ensures the safety of other AI fast takeoffs. Dewey also suggests that two types of strategic advantage, proactive or reactive, may be reached by non-self-improving AI. In his opinion, another option is strategic advantage reached by non-AI technological means.

4.1.3. Risks of Creating Hard-takeoff AI as a Global Solution

In AI safety research, it is often assumed that the first superintelligent AI will take action to prevent the creation of other AIs. In that case, solving local AI safety would provide global safety.

However, if the first AI is created in, say, the US, it must then prevent the creation of another AI in, say, China. From the point of view of international law, such an action by an AI could be an act of war [40].

Deliberately creating an AI that will start a war immediately after its creation is very provocative for other actors. In the face of such a threat they might use a preemptive nuclear strike to prevent the creation of AI. Kahn [48] wrote the same of the potential creation of a Doomsday nuclear bomb that could kill all humanity—that just the act of its creation could be even more provocative than a nuclear attack.

Not just the actual creation, but just the intention to create such AI, may attract attention from foreign and domestic secret services. Publicly suggesting that the first creators of AI should program it to take over the world may have legal consequences (as such an AI could be classified as a cyberweapon) and may prevent open dissemination of any AI safety theory based on such a suggestion.

It appears that creation of a military infrastructure is a convergent instrumental goal for any first AI [40]. This infrastructure would help the AI prevent the creation of other AIs as well as prevent humans and government agencies from trying to switch off the AI. If other AIs are in advanced stages of development, they will resist the attempt to shut them down. In this case, a war between AIs will start, in which humanity could perish or be taken hostage. Therefore, this solution is intrinsically risky and better solutions should be sought.

Another idea is that the creation of AI safety theory will happen separately from the creation of AI, but the first AI creator will use available safety theory. We will discuss this possibility below.

4.2. One Global AI Created by Collective Efforts

4.2.1. AI Nanny Requires a World Government for Its Creation

The idea of an AI Nanny has been suggested by Ben Goertzel, who has described " . . . the creation of a powerful yet limited Artificial General Intelligence (AGI) system . . . with the explicit goal of keeping things on the planet under control while we figure out the hard problem of how to create a probably positive Singularity. That is: to create an 'AI Nanny'" [21]. He proposed the following properties for an AI Nanny:

- General intelligence somewhat above the human level,
- Interconnection with powerful worldwide surveillance systems,
- Control of a massive contingent of robots, and
- A cognitive architecture featuring an explicit set of goals.

Muehlhauser and Salamon [49] criticized this idea because solving AI safety for the AI Nanny would require solving almost all AI safety problems for self-improving AI.

The AI Nanny also does not solve the main problem of how the first AI will gain its global power—by world takeover or by peaceful integration of a net of AIs. The first way has its own risks and the second could have dangerous holes. One possible solution here is peaceful integration of most of the world, and the forceful integration of any remaining "rogue states." This could resemble the current dynamic between a large international coalition of nuclear-armed states with "rogue countries" that try to make their own nuclear weapons.

A united world government may be required for the creation of an AI Nanny, but under current conditions, such a world government is unlikely to peacefully appear. Such a world government might appear if one country gained an overwhelming military advantage from a means other than AI. If the advantage arose from AI, the problem of AI safety would already be solved, but it could come from powerful nanotechnological weapons or some type of narrow-AI robotics. Alternatively, if the risks of AI are highly visible, or perhaps already felt, most countries may give up their sovereignty to the UN to create an AI Nanny. Such a scenario could happen if a narrow-AI-based computer virus created widespread devastation of infrastructure, or if the first self-improving AI appeared, but spectacularly failed at some stage of its development.

The AI Nanny may have rather high intelligence, but in a form which is not easy to self-improve, e.g., a large database of pre-recorded solutions and neural algorithms, as well as all existing data about the world and new data from surveillance systems. Such a "data-driven" AI may be a relatively safe local solution.

Some semi-universal AI may be created in the current age of neural nets [50] as a very large and prohibitively expensive international project, for example, the Human Genome Project, Large Hadron Collider, and International Thermonuclear Experimental Reactor. Gary Marcus recently suggested that we need something analogous to the European Organization for Nuclear Research, CERN, for AI [51], in a sense similar to Baruch's 1946 plan to centralize nuclear research [10].

An AI Nanny could be designed on many opaque neural net modules that would prevent its self-improvement, and its enormous size would prevent it from leaking into the internet. Its intelligence also may not be universal or not exceed total human intelligence. Therefore, an AI Nanny would likely be rather safe and under international control. However, the opportunity for such a project may be lost, as many large companies are now participating in their own projects and there is a lot of available hardware as well as openly published materials. Yet the potential is not completely lost; large international collaborations such as the "Partnership on AI" [52] could contribute momentum to the creation of an AI Nanny, if they chose to do so.

4.2.2. Levels of Implementation of the AI Nanny Concept

We suggest four levels of possible intelligence of an AI Nanny:

1. Use of a distributed surveillance system, which does not have much intelligence but is able to enforce a universal ban on creation of self-improving systems. This is a low-level solution.

2. Creation of neural-net-based and data-driven AI as part of a large international project. In this case, the AI's intelligence comes not from fluid intelligence but from extensive knowledge and models. It may serve as the brain of the surveillance system mentioned above. One possible solution could be to use an upload human as an "AI king," or world governor, with the main mission of preventing the creation of other AIs [53]. Such an AI king would run at higher speeds than ordinary humans, using all available hardware, which will give it greater intelligence while maintaining alignment with human values. This idea would be obviously controversial from technical, political, and moral points of view.

3. Creation of AI police, a net of narrow AIs able to control the appearance of self-improving AIs and other dangerous entities.

4. Creation of a high-intelligence AI Nanny as described by Reference [21]. This AI would be some form of superintelligence (SI), as much above humans as humans are above apes. In this case,

there would be exactly the same problems as with the control of any other strong AI [49]. However, if the system were weaker, it might be possible to find Goldilocks' path between its ability to control research and our ability to control the system.

4.2.3. Global Transition into AI: Non-Agential AI-Medium Everywhere, Accelerating Smoothly without Tipping Points

The AI described above was agential. However, some of the strongest known optimization processes are non-agential: e.g., evolution, market forces, and science. These processes appear from the interaction of millions of agents with their own goals, and the optimization power of these processes does not depend much on direct summing of the minds of agents. Instead, it is a result of their interactions, so it is not a net of AIs, which will be discussed below, as a net implies higher level of goal's coordination.

The AI-medium self-improves more quickly than any individual part of it, because self-improvement is a property of the whole system, but not of any one part of it, as it results from the way information is exchanged between different parts.

We will call such processes "intelligent media," as opposed to intelligent agents, as they do not have independent goals, but perform any tasks they find. This medium is a form of environment; as such, it does not conquer territories, but attracts other agents to participate in it; a similar idea has been suggested by Mahoney [54]. This feature could still be devastating, as we know that in an analogous case, market forces can destroy traditional cultures more effectively than weapons [55]. A non-agential AI-medium does not have to take over the world because it would simultaneously appear everywhere.

It would not be surprising if superintelligence also arises from a medium. This idea in naïve form has been presented as "the internet will gain consciousness." The internet surely will be a backbone for the AI-medium, but something more is needed. One can imagine other elements of an AI-medium in the form of blockchain, social networks, prediction markets [56], and the network of scientific references [57]. One of the routes to an AI-medium could be to connect all human brains through some form of network, producing, in effect, a global brain [58].

There are concerns that such collective evolution is unstable and will eventually produce one agent that will be able to improve itself more quickly than the overall AI-medium and thus destroy it. See, for example, Sotala's review criticizing Vinding's recent book discussing the difference between individual and collective takeover [9,59].

Scott Alexander argues that an accelerating self-improving AI-medium is possibly a negative outcome as it could take the form of an "ascending economy" [55], where a group of market agents create an evolving ecosystem, which destroys all human values in order to increase "growth." Karl Marx criticized market economics for the same flaw [60].

As the AI-medium naturally evolves without taking into account human values, it cannot be considered friendly or unfriendly to humans. Given this, unless regulated in some way, it will either prioritize human survival, if humans will be able to positively interact with it, or it will ignore humans. John Smart [61] predicted that the evolution of such a system will consist of constant acceleration and miniaturization, which could be described by a hyperbolic law.

Drexler suggested another form of AI-medium, Comprehensive AI Services [14], in which superintelligence does not have agency. Instead, it consists of many narrow superintelligent tools, which also could be used to create needed level of surveillance to prevent rogue AI appearance elsewhere. A primitive example of such service now is Google with its many "Tool AIs": web search, email, drive, which are integrated in one ecosystem but are not agential. However, as Gwern wrote [62], any Tool AI "wants" to be agential AI, as it would increase its efficiency, and thus AI Services could eventually turn into or spawn potentially dangerous agential AI.

4.3. Help Others to Create Safe AI

4.3.1. Promoting Ideas of AI Safety in General and the Best AI Safety Solution to All Players

Helping others develop improved AI safety is a global solution if there are ways to reach all significant AI players.

As we mentioned above, it is a priori improbable that the same team that creates an optimal AI safety theory will also create the first AI unless it is part of an international collaboration. Therefore, teams working on AI safety should try to convince other AI teams to adopt the best AI safety theory.

There are a number of tangential measures that may help in the development of AI safety, but do not guarantee good results, including:

- Funding of AI safety research.
- Promotion of the idea of AI safety.
- Protesting military AI.
- Friendly AI training for AI researchers.
- Providing publicly available safety recommendations.
- Increasing the "sanity waterline" and rationality in the general population and among AI researchers and policymakers.
- Lowering global levels of confrontation and enmity.
- Forming political parties for the prevention of existential risks and control of AI risks, or lobbying current political parties to adopt these positions. However, even if such parties were to win in larger countries and were able to change policy, there would still be countries that could use any technology "freeze" in larger countries to their advantage.

Another idea is to seek ways to attract the best minds to solve the AI safety problem. Yudkowsky said that one of reasons he wrote the book 'Harry Potter and the Methods of Rationality' [63] was to attract the best mathematical minds to the AI safety problem. Attracting top minds would achieve simultaneously several useful goals:

- Depleting the pool of minds for direct—not necessarily safe—AI research, thereby slowing it down
- Increasing the quantity and quality of thought working on AI safety theory
- Establishing relationships between the best AI teams, as some of the people who will have worked on AI safety may have come from such teams, may eventually join them, or may otherwise have friends there, and
- Promoting the idea that unlimited self-improvement is dangerous and unstable for all players, including AIs.

4.3.2. Selling AI Safety Theory as an Effective Tool to Align Arbitrary AI

One possible way to reach many people is to make the solution attractive. If AI safety implementation can be used to align the goals of an arbitrary AI, it will be very attractive for any reasonable AI creator, as the creator insures their own safety and ability to place goals into the AI. The AI creator could save many resources by implementing a proven alignment method. However, while this lessens the probability that the AI will run amok, the creator could still align the AI with a dangerous, egoistic goal.

If an AI safety tool-kit could be sold as a good, this would increase the likelihood that first movers will use it, as it would be profitable for them. It could also be sold as a service, which could include custom adaptation and training. Selling "AI safety" may produce a wider reach than just publishing a PDF with explanations, and the customer support could increase its implementability.

4.4. Local Action to Affect Other AIs Globally

4.4.1. Slowing the Appearance of Other AIs

In this scenario people could take actions locally that will affect any other AI globally, which may appear in the future at an unknown location.

Such actions may include espionage or taking low-hanging fruit in research, which will increase overall level of the technology, but lower chances that one of the participants of the race will leapfrog others by taking such low-hanging fruit; draining the pool of easily available resources, which includes both minds and hardware, may also be regarded as taking low-hanging fruits. While it is impossible to drain all hardware, the leader in AI research could invest in owning leading positions in hardware capabilities as well as training datasets for neural nets.

4.4.2. Ways to Affect a Race to Create the First AI

An AI creation race is generally regarded as bad because it encourages the creation of the least-safe AIs first. A war between AIs may also become possible if several AIs are created simultaneously [64,65].

There are many ideas on how to affect an AI race in order to make it safer, that is, to lower the probability of creating dangerous AI. As a race with many participants is a very complex game, there are not obvious ways to predict how it will react to seemingly good interventions, for example, openness. Bostrom has shown that if no one knows the capabilities of others and their own capabilities, it will slow down the race, so openness about capabilities may be dangerous [66].

Actions that may affect an AI race and make it safer may include:

- Changing the number of participants.
- Increasing or decreasing information exchange and level of openness.
- Reducing the level of enmity between organizations and countries, and preventing conventional arms races and military buildups.
- Increasing the level of cooperation, coordination, and acceptance of the idea of AI safety among AI researchers.
- Changing the total amount of funding available.
- Promoting intrinsic motivations for safety. Seth Baum discussed the weakness of monetary incentives for beneficial AI designs, and cautions: "One recurrent finding is that monetary incentives can reduce intrinsic motivation" [67]; when the money is gone, people lose motivation. Baum also noted that the mere fact that a law existed promoted obedience in some situations and that social encouragement can increase intrinsic motivation.
- Changing social attitudes toward the problem and increasing awareness of the idea of AI safety.
- Trying to affect the speed of the AI race, either slowing it down or accelerating it in just one place by concentrating research. It is interesting to note that acceleration could be done locally, but slowing it would require global cooperation, and so is less probable.
- Affecting the idea of the AI race as it is understood by the participants [67]: if everybody thinks that the winner takes everything, the race is more dangerous. A similar framing solution has been suggested in the field of bioweapons, that is, to stop claiming bioweapon creation is easy, as it might become attractive to potential bioterrorists. In fact, bioweapons are not as easy to develop and deploy as is shown in movies, and would probably kill the terrorists first [68].
- Affecting the public image of AI researchers who are currently presented as not wanting beneficial AI design [67].
- Refraining from suggestions of draconian surveillance as they "inadvertently frame efforts to promote beneficial AI as being the problem, not the solution" [67].
- Stigmatization of building recursive self-improving AI by framing them as morally unacceptable, as has been done with landmines. The stigma impelled even countries that did not sign the treaty that prohibits landmines to reduce production [67].

- Deliberate association with crackpottery: an example is UFO (Unidentified Flying Objects) research: anyone who mentions the word "UFO" will no longer be accepted in the scientific community as a credible scientist. This partially worked against AI during past AI winters, when scientists tried not to mention the words "artificial intelligence." Society could come to associate "self-improving AI" with craziness, which would be not difficult if we pick some of the most outstanding ideas from associated internet forums, e.g., Roko's Basilisk [69]. Such an association may reduce funding for such research. However, AI could start to self-improve even if it was not designed to do so; thus, such association would probably be damaging to AI safety efforts. Recent successes in meta-learning in neural nets by DeepMind show that the idea of self-improving AI is becoming mainstream [70].
- Affecting the speed of takeoff after one AI starts to win. If the speed of self-improvement of one AI diminishes, other AIs may catch up with it.

We address some of these ideas in the next section.

4.4.3. Participating in Acausal Deals with Future AI

Rolf Nelson [71] suggested that we could install indexical uncertainty into the future AI; in that case, if we make a commitment now that if humanity creates a friendly AI, this friendly AI will also create simulations of most probable types of rogue AI, which will be turned off if a given AI does not simulate benevolence to humans. In that case, any rogue AI will be uncertain if it is in a simulation or not, and as killing humans has small marginal utility in most cases, it would prefer to display benevolence. However, such an approach would probably work only for an AI singleton, and it is our last level of defense.

5. "Many AI" Solutions

5.1. Overview of the "Net Solutions" of AI Safety

5.1.1. How a Net of AIs May Provide Global Safety

In a nutshell, the idea of a "net solution" to AI safety is that there will be many AIs, and this fact will provide some form of protection. The most prominent backer of this approach is Elon Musk, who wants to unite AI working teams in a net based on openness and upgrade humans, so they will not become obsolete in the age of AI [72]. However, there are risks to this approach [66].

There are two main features, which may provide safety with a net of AIs:

1. The combined intelligence of many AIs (the net of AIs) is much higher than the one of any rogue AI, so the net is able to create effective protection. An AI-net could form something similar to AI "police," which prevent any single AI from unlimited growth. This is analogous to the way the human body provides a multilevel defense against unlimited growth of a single cancerous cell in the form of an immune system. The approach is somewhat similar to the AI Nanny approach [21], but an AI Nanny is a single AI entity. An AI-net consists of many AIs, which use ubiquitous transparency [73] to control and balance [74] each other.
2. Value diversity among many AI-sovereigns [2,75] guarantees that different positive values will not be lost. Different members of the net have different terminal values, thus ensuring diversity of values, as long as the values do not destructively interfere. If the values do destructively interfere, then solutions must be found for these conflicts.

We will call this many AI solution a "Multipolar Singleton" [11], as global coordination will result from constant negotiation and trade between entities with different values. A Multipolar Singleton will have the following necessary conditions:

- Many superhuman AIs exist.
- The AIs all find mutual cooperation beneficial, and have some mechanism for peaceful conflict resolution.
- The AIs have diversity of final goals, so some goals are more beneficial to humans than others. This protects against any critical mistake in defining a final goal, as many goals exist. However, it is not optimal, as some of AIs may have goals that are detrimental for humans. It will be similar to our current world, with different countries, but the main difference will be that they will likely be much better able to peacefully coexist than currently, because of AI support.
- Because Earth is surrounded by infinite space, different AIs could start to travel to the stars in different directions, and as each direction includes a very large number of stars, even very ambitious goals could be not mutually exclusive and might not provoke conflicts and wars.
- Finding it mutually beneficial to create AI police to prevent unlimited self-improving AIs or other dangerous AIs from developing via ubiquitous intelligent control.

The main question is how to reach an AI-net solution and whether it will be stable, collapse into war between AIs, or reduce to a single AI dictatorship.

5.1.2. The Importance of Number in the Net of AIs

The most important variable here is the number of future superintelligent AIs, which depends on the speed of AI self-improvement and the number of teams of AIs creators, as well as the upper limit of individual intelligence, if it exists. The slower the AI takeoff is, the larger the number of AIs will be, though this also depends on the number of AI teams, among other factors. There are several vague groups of the possible numbers of coexisting superintelligences, which will have different dynamics, including:

- Two AI-sovereigns' semi-stable solution, similar to the Cold War [76].
- From several to dozens of sovereign AIs, similar to existing nation-states; they may be evolved from nation-states, or from large companies.
- From thousands to billions of AIs, with relations similar to relations between humans now, possibly resulting from some brain uploading technology [74], human augmentation [77], or genetic modification [78], but each single AI is not significantly above human level.
- Uncountable or almost-infinite number of AIs, similar to AI-medium, discussed above. This could be similar to the IoT, but with AIs as nodes.

In the following list we present an overview of possible solutions, which will be explored in detail below.

- Net of AIs forms a multilevel immune system to protect against rogue AIs and has a diversity of values, thus including human-positive values
    - Instruments to increase the number and diversity of AIs:
        - openness
        - slowdown of AI growth
        - human augmentation
        - self-improving organizations
        - increase number of AI teams
        - create many copies of the first AIs
    - Net of AIs is based on human uploaded minds

- Several AI-sovereigns coexist, and they have better defensive than offensive capabilities
    - ○ Two AIs semi-stable "Cold war" solution, characterized by
        - tight arms race
        - military AI evolution
        - MAD defense posture
    - ○ AI-sovereigns appear from nation-states
        - very slow takeoff and integration with governments
    - ○ Different AIs expand in space in different directions without conflict
    - ○ Creation of AIs on remote planets

5.2. From the Arms Race between AI-Creating Teams to the Net of AIs

5.2.1. Openness in AI Development

Elon Musk and others presented the idea of OpenAI in 2015: "We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as is possible safely" [29]. In the following, we discuss the idea of openness of the field of AI and the net of AIs as we understand it; it does not represent the position of the "OpenAI" initiative. We look at the following approach: many AI projects freely exchange ideas, datasets and progress results, thus accelerating AI creation and ensuring its safety. We will call it an "open net of AI teams".

Safety emerges from the following characteristics of such collaboration:

- None of the AI teams gains strategic advantage over other teams, as all the data from every team's results are available to all of the teams. An attempt to hide results will be seen publicly. Openness ensures that many AI teams will come close to self-improving AI simultaneously, and that there will be many such AIs, which will balance each other.
- The teams outside the "open net" are much less likely to gain strategic advantage, as they are not getting all the benefits of the membership in the net, namely access to the results and capabilities of others. However, this depends on how much information becomes part of the public domain. The open net will have an "intelligence advantage" over any smaller player, which makes it more probable that self-improvement will start inside the open net, or that the net will have time to react before a rogue agent "outsmarts" the net.
- The "open net" will create many different AIs, which will balance each other, and probably will be motivated to engage in mutually useful collaboration (However, some could take advantage of openness of others but not share their own data and ideas). If one AI leaves the net for uncontrolled self-improvement, the collective intelligence of the net will still be higher than that AI for some time, probably enough to stop the rogue AI.
- The value system of the net will provide necessary diversity, so many possible goals will be presented to at least some extent. This lessens the chance that any good goal will lost, but raises the chance that some AI projects will have bad, dangerous, or otherwise unacceptable goals.
- Because of their ability to collaborate, the "open net" may be able to come to unanimous decisions about important topics, thus effectively forming a Singleton.
- The net will be able to assess and possibly control all the low-hanging fruits of self-improvement, for example, the ability to buy hardware or take over the internet, thus slowing down self-improvement of any rogue agent.
- The net will help to observe what the other players, not involved in open net, are doing—for example, the fact that German scientists stopped publishing articles about uranium in 1939 showed that they were trying to keep their work secret and therefore indirectly hinted that they were working on a bomb.

- The net will contribute to the creation AI vigilantes or AI police, as suggested by David Brin in his transparent society proposal [73]. Therefore, the open net may somehow evolve in the direction of an AI Nanny, perhaps consisting of many distributed nodes.

Bostrom has criticized the idea of openness in AI, because he feels it could accelerate dangerous research [66]. It would also not be easy to balance a dangerous AI, as it could undertake local actions that could quickly kill everybody, like constructing a very large nuclear cobalt bomb [79] or a dangerous biological virus. However, if there are many AIs, they probably could have the needed level of mutual control to prevent local dangerous actions or contain the results of such actions.

The main question is if openness in AI will be able to prevent a rogue actor from using these data to start self-improving first and gaining a decisive advantage over others. These worries are described in an excellent post by Scott Alexander [80].

Above we assumed that the net of teams will create the net of AI, however, the net of teams may cooperate in creating just one AI.

5.2.2. Increase of the Number of AI Labs, so Many AIs Will Appear Simultaneously

Bostrom explored the situation of many competing teams depending of their number, their enmity, and their knowledge about their own and each others' capabilities. He found that the fewer the number of the teams, the smaller the overall risk, and also that it is better if they do not know about each other's or their own capabilities [64].

In fact, there are already many AI teams and such a large number may result in many simultaneous AI takeoffs. History shows that some important discoveries were made independently with a very small temporal separation. For example, the first two telephone patent applications were filed within three hours of each other on 14 February 1876 [81] and the Soviet–US race to bring material back to Earth from the Moon was decided by three days in 1969 [82].

Increasing the number of independent AI projects will increase the probability that several of them will have hard takeoffs simultaneously, but it will also increase the chances that some of the programs will have a very low level of safety, as Bostrom et al. note [64].

The publicity around AI in recent years has likely contributed to the growth of AI companies. Venture Scanner tracked 957 AI-creating companies [83]. While most of them are not trying to build AGI, many of them would be happy to have an AI as universal as possible. It is also clear that many companies and individuals are not presented in this list, including university projects and individual researchers. It could also be that some companies on the list are fake or should not be counted for other reasons. Therefore, it is reasonable to estimate the total number AI teams now working as within an order of magnitude of 100, but most of the research is coming from around ten major companies including Google, Facebook, and Open AI.

This means that there may be no need to increase the number of teams to prevent a single dominant AI—their number is already on the order of magnitude where several hard takeoffs could happen simultaneously.

5.2.3. Change of the Self-Improving Curve Form, So That the Distance Between Self-Improving AIs Will Diminish

Yampolskiy has argued that there are several reasons why the actual self-improving of one AI system may be described by a logarithmic rather than exponential curve [84]. However, artificial interventions such as taking low-hanging fruits or espionage could change this rate. If the curve is shallower, more AIs will reach the level of superintelligence simultaneously, providing a better chance for some balance of power.

5.3. Instruments to Make the Net of AIs Safer

5.3.1. Selling Cheap and Safe "Robotic Brains" Based on Non-Self-Improving Human-Like AI

This idea is to make a safe AI design, which can solve almost all tasks that other people or organizations may need. Such a design would then be provided widely and very cheaply either as hardware or from the cloud. This would undermine the economic need for creation of other AIs and create the opportunity for a global AI Nanny. This non-self-improving, safer AI is analogous to the idea of non-self-replicating safer molecular manufacturing, such as a nanofab, which is regarded a safer form of nanotech than nanorobots [85].

One possible design of such a "robotic brain" could be a human upload [74] or some simplified model of a human brain, which finds a balance between upload and neuromorphic AI [53].

5.3.2. Starting Many AIs Simultaneously

Any AI-creating team could start not one, but many AIs, just to balance the possible flows in the first AI or to observe its possible flaws depending on initial conditions of different AI. Such an approach will likely have unpredictable consequences, and might be used only as a backup measure, if control over the first seed AI is lost. This is applicable to any AI—if the control over it has been lost, another copy of the same AI could be started from the backup with slight changes of goal function. However, the idea of "beneficial computer viruses" was already discussed and it was concluded that such viruses would not be better than normal antivirus software, as the second virus, intended to deactivate the first one, will spread much less and will also cause harm [86].

6. Solutions in Which Humans are Part of the AI System

6.1. Different Ways to Incorporate Humans inside AI

Some form of superintelligence may be created with humans as participants within it, as Drexler suggested in his Comprehensive AI Services. However, as Bostrom shows [2], there is always the problem of the "second transition," that is, the appearance of a more powerful AI inside such a system, one which no longer needs humans. So, any such system would need to create an AI police to prevent a "second transition."

Another problem is that most such solutions are lagging, as human uploading is technically still far away, if possible at all.

Some ways of incorporating humans inside AI include:

- AI could be built around a human core or as a human emulation. It could result from effective personal self-improvement via neural implants [2], adding tool AIs and exocortex. There is no problem of "AI alignment," as there are not two agents that should be aligned, but only one agent whose value system is evolving [53], however, if the human core is not aligned with the rest of humanity, the same misalignment problem could appear—therefore the ethics of the core human are crucial.
- AI could appear from a net of self-improving posthumans, connected via neural interfaces [87]. This combines ideas of social networks, blockchain, and Neuralink [77]. Such a net could conceivably appear from the evolution of some types of medical AI [88].
- AI could result from genetic modification of humans for intelligence improvement [78].
- Superintelligence could appear as a swarm intelligence of many human uploads and not evolve in a more effective and less human form for some unknown reason [74,89].
- Only one human upload is created, and it works as an AI Nanny, preventing the emergence of any other superintelligences [53].
- Superintelligence is created by a "self-improving organization" as a property of the whole organization, which includes employees, owners, computers, hardware-building capabilities,

social mechanisms, and owners. It could be a net of self-improving organizations, similar to Open AI [29] or the "Partnership on AI".

- Nation-states evolve into AI-states, and keep most of their legislation, structure, values, people, and territories. This is most probable in the case of the soft takeoff scenarios, which would take years. Earth could evolve into a bipolar world, similar to the Cold War, or a multipolar world. In this scenario, we could expect a merger between self-improving organizations and AI-states, perhaps by acquisition of such companies by state players.

Is not easy to envision them at this point, but there could also be scenarios which combine some of the ideas in this section.

6.2. Even Unfriendly AI Will Preserve Some Humans or Information about Humans

Below is an assortment of less-probable ideas that generally provide a lower level of safety (Levels 1 and 2). In these scenarios, human beings will somehow be incorporated, used, or remembered by unfriendly AI.

- Unfriendly AI may have a subgoal to behave as benevolent AI toward humans, based on some Pascal mugging-style considerations and ontological uncertainty if it will think that there is small chance that it is in a simulation which tests its behavior [71].
- Even unaligned AI will probably model humans in instrumental simulations [90] needed to solve the Fermi paradox.
- Humans could be cost-effective workers in some domains and might therefore be retained, though only to be treated as slaves.
- AI could preserve some humans as a potentially valuable asset, perhaps to trade information about them with potential alien AI [75], or to sell them to a benevolent AI.
- AI may still preserve information about human history and DNA for billions of years, even if the AI does not use or simulate humans in the near term. It may later return them to life if it needs humans for some instrumental goal.
- AI may use human "wetware" (biological brains) as efficient supercomputers.
- AI could ignore humans and choose to live in space, while humans would survive on Earth. AI would preserve humanity if the marginal utility derivable from humanity's atoms is less than the marginal instrumental utility from humanity's continued existence.

As human values are formed by evolution, an evolving AI system [61] may naturally converge to a similar set of values as humans [91].

7. Which Local Solutions Are the Best to Get a Stable Global Solution?

In the sections above, we overviewed all (to the best of our knowledge) previously suggested global solutions for AI safety.

There are many possible global solutions to the AI safety problem, but humanity must choose the one that has the highest probability of successful implementation.

Clearly, a "no AI" solution should not be implemented, as it would be unethical (due to opportunity costs) and ineffective (due to intense pressures to achieve AI). As "AI safety theory" is lagging current AI development; a controllable, self-improving AI as a global solution will probably not be possible in the next couple of decades. We also lack the global coordination [92] to create an AI Nanny, as well as the technologies necessary for human uploading.

Neural-net-based solutions developed by major IT companies are currently the greatest technological source of success in AI research [70,93,94]. Such organizations not only create AI, but improve their own organizational structure by similar processes, giving rise to "self-improving organizations." Google ("Alphabet") is the leader here by a large margin.

Soft acceleration of several self-improving organizations seems to be the most plausible way to build a mild form of superintelligence in the current epoch, a plan Christiano named "prosaic AI" [50]. It may also be fueled by an AI race between the US and China [95].

In the current technological and political situation, several local approaches seem to be most safely scalable to the global scale:

(1) Comprehensive AI Services, which could become a basis for a system of ubiquitous surveillance and AI Police, preventing appearance of rogue AIs.
(2) Research in human uploads or human-mind models, which will result in many AIs of relatively limited capabilities [74]. This again could be used to create AI Police.
(3) Self-improving organizations, where humans and AI work together which is basically is part of Drexler's suggestion [14], but also could be done in Christiano's approach of iterated amplification and factored cognition [96].
(4) Robotic mind-bricks, that are pre-trained AI with limited capabilities and prefabricated safety measures which would be sold widely and provide a basis for global AI policing.
(5) AI Safety as a service, similar in some sense to current antivirus computer industry.

8. Conclusions

Suggested solutions to AI safety problem are either local or global, and when choosing a local solution, we also should take into account how it could be safely scaled globally. In this article, we posed the problem of relation between global and local solutions, overviewed existing global solutions and estimated their safety.

We identified a group of local solutions which seems to be more easily and safely scaled into a global level. This group includes such approaches as Comprehensive AI Services, selling robotic mind-bricks, and AI Safety as a service.

9. Disclaimer

This article represents views of the authors and does not necessarily represent the views of the ALLFED, the Markkula Center for Applied Ethics, or other organizations to which the authors belong.

Author Contributions: A.T. provided writing—original draft preparation, D.D. provided writing—review and editing, and B.P.G. provided writing—review and editing.

Funding: This research received no external funding.

Acknowledgments: We thank Anthony Barrett for insightful comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yampolsky, R.; Fox, J. Safety engineering for artificial general intelligence. *Topoi* **2013**, *32*, 217–226. [CrossRef]
2. Bostrom, N. *Superintelligence*; Oxford University Press: Oxford, UK, 2014.
3. Russell, S. 3 Principles for Creating Safer AI. Available online: https://www.youtube.com/watch?v=EBK-a94IFHY (accessed on 18 February 2019).
4. Sotala, K.; Yampolskiy, R. Responses to catastrophic AGI risk: A survey. *Phys. Scr.* **2015**, *90*, 069501. [CrossRef]
5. Yudkowsky, E. *Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks*; Cirkovic, M.M., Bostrom, N., Eds.; Oxford University Press: Oxford, UK, 2008.
6. Christiano, P. Takeoff Speeds. Available online: https://sideways-view.com/2018/02/24/takeoff-speeds/ (accessed on 5 March 2018).
7. Ramamoorthy, A.; Yampolskiy, R. Beyond MAD?: The race for artificial general intelligence. *ICT Discov. Spec. Issue* **2018**, *1*, 1–8.

8. Brundage, M.; Avin, S.; Clark, J.; Toner, H.; Eckersley, P.; Garfinkel, B.; Dafoe, A.; Scharre, P.; Zeitzoff, T.; Filar, B. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv* **2018**, arXiv:1802.07228.
9. Sotala, K. Disjunctive AI Scenarios: Individual or Collective Takeoff? 2017. Available online: https://kajsotala.fi/2017/01/disjunctive-ai-scenarios-individual-or-collective-takeoff/ (accessed on 18 February 2019).
10. Dewey, D. *Long-Term Strategies for Ending Existential Risk from Fast Takeoff*; Taylor & Francis: New York, NY, USA, 2016.
11. Bostrom, N. What is a singleton. *Linguist. Philos. Investig.* **2006**, *5*, 48–54.
12. Krakovna, V. Risks from general artificial intelligence without an intelligence explosion. *Deep Saf.* **2015**, *26*, 1–8.
13. Turchin, A.; Denkenberger, D. Classification of Global Catastrophic Risks Connected with Artificial intelligence. *J. Br. Interpanet. Soc.* **2018**, *71*, 71–79. [CrossRef]
14. Drexler, K.E. Reframing Superintelligence. Available online: https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf (accessed on 18 February 2019).
15. Turchin, A. Assessing the future plausibility of catastrophically dangerous AI. *Futures* **2018**. [CrossRef]
16. Beckstead, N. *On the Overwhelming Importance of Shaping the Far Future*; Department of Philosophy, Rutgers University: New Brunswick, NJ, USA, 2013.
17. Bostrom, N. Existential risks: Analyzing Human Extinction Scenarios and Related Hazards. *J. Evol. Technol.* **2002**, *9*, 2002.
18. Torres, P. Problems with Defining an Existential Risk. Available online: https://ieet.org/index.php/IEET2/more/torres20150121 (accessed on 18 February 2019).
19. Green, B.P. The Technology of Holiness: A Response to Hava Tirosh-Samuelson. *Theol. Sci.* **2018**, *16*, 223–228. [CrossRef]
20. Tomasik, B. *Artificial Intelligence and Its Implications for Future Suffering*; Foundational Research Institute: Basel, Switzerland, 2017.
21. Goertzel, B. Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? *J. Conscious. Stud.* **2012**, *19*, 96–111.
22. Yudkowsky, E. Coherent Extrapolated Volition. Available online: http://intelligence.org/files/CEV.pdf (accessed on 18 February 2019).
23. Weng, Y.-H.; Chen, C.-H.; Sun, C.-T. Safety Intelligence and Legal Machine Language: Do We Need the Three Laws of Robotics. In *Service Robot Applications*; InTech: Rijeka, Croatia, 2008.
24. Hughes, J. Relinquishment or Regulation: Dealing with Apocalyptic Technological Threats. *Hartford CT Novemb.* **2001**, *14*, 06106.
25. Yudkowsky, E. *There's No Fire Alarm for Artificial General Intelligence*; Machine Intelligence Research Institute: Berkeley, CA, USA, 2017.
26. Robots: Legal Affairs Committee Calls for EU-Wide Rules. Available online: http://www.europarl.europa.eu/news/en/press-room/20170110IPR57613/robots-legal-affairs-committee-calls-for-eu-wide-rules (accessed on 18 February 2019).
27. Future of Life Institute Asilomar AI Principles. Available online: https://futureoflife.org/ai-principles/ (accessed on 18 February 2019).
28. Morris, D.Z. Elon Musk: Artificial Intelligence Is the "Greatest Risk We Face as a Civilization". Available online: http://fortune.com/2017/07/15/elon-musk-artificial-intelligence-2/ (accessed on 18 July 2017).
29. Brockman, G.; Sutskever, I. Introducing OpenAI. Available online: https://openai.com/blog/introducing-openai/ (accessed on 18 February 2019).
30. Berglas, A. Artificial intelligence will kill our grandchildren (singularity). Unpublished work, 2012.
31. Green, B. Are science, technology, and engineering now the most important subjects for ethics? Our need to respond. In Proceedings of the 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, Chicago, IL, USA, 23–24 May 2014; pp. 1–7.
32. Green, B. Emerging technologies, catastrophic risks, and ethics: three strategies for reducing risk. In Proceedings of the 2016 IEEE International Symposium on Ethics in Engineering, Science and Technology (ETHICS), Vancouver, BC, Canada, 13–14 May 2016.
33. List of Semiconductor Fabrication Plants. Available online: https://en.wikipedia.org/wiki/List_of_semiconductor_fabrication_plants (accessed on 18 February 2019).

34. Cole, D.D.; Denkenberger, D.; Griswold, M.; Abdelkhaliq, M.; Pearce, J. Feeding Everyone if Industry is Disabled. In Proceedings of the 6th International Disaster and Risk Conference, Davos, Switzerland, 28 August–1 September 2016.
35. Denkenberger, D.; Cole, D.; Griswold, M.; Pearce, J.; Taylor, A.R. Non Food Needs if Industry is Disabled. In Proceedings of the Proceedings of the 6th International Disaster and Risk Conference, Davos, Switzerland, 28 August–1 September 2016.
36. Jones, S.E. *Against Technology: From the Luddites to Neo-Luddism*; Routledge: Abingdon, UK, 2013; ISBN 1-135-52239-1.
37. Kushner, D. The real story of stuxnet. *IEEE Spectr.* **2013**, *50*, 48–53. [CrossRef]
38. Bostrom, N. The Unilateralist's Curse: The Case for a Principle of Conformity. Available online: http://www.nickbostrom.com/papers/unilateralist.pdf (accessed on 18 February 2019).
39. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Knopf: New York, NY, USA, 2017.
40. Turchin, A.; Denkenberger, D. Military AI as convergent goal of the self-improving AI. In *Artificial Intelligence Safety and Security*; CRC Press: Louisville, KY, USA, 2018.
41. Teller, E. *LA-602: The Ignition of Atmosphere with Nuclear Bombs*; Los Alamos Laboratory: Los Alamos, NM, USA, 1946.
42. Ria Novosti Испытания ядерного оружия на Тоцком полигоне. Справка. Available online: https://ria.ru/defense_safety/20090914/184923659.html (accessed on 18 July 2017).
43. Nuclearweaponarchive India's Nuclear Weapons Program—Smiling Buddha: 1974. Available online: http://nuclearweaponarchive.org/India/IndiaSmiling.html (accessed on 18 July 2017).
44. MIRI. MIRI AMA—Anyone May Ask. Available online: http://effective-altruism.com/r/main/ea/12r/ask_miri_anything_ama/ (accessed on 20 February 2019).
45. MIRI. About MIRI. Available online: https://intelligence.org/about/ (accessed on 18 February 2019).
46. Sotala, K. Decisive Strategic Advantage without a Hard Takeoff. 2016. Available online: https://kajsotala.fi/2016/04/decisive-strategic-advantage-without-a-hard-takeoff/ (accessed on 18 February 2019).
47. Putin, V. Open Lesson "Russia Looking to the Future". Available online: http://kremlin.ru/events/president/news/55493 (accessed on 28 October 2017).
48. Kahn, H. *On Thermonuclear War*; Princeton University Press: Princeton, NJ, USA, 1959.
49. Muehlhauser, L.; Salamon, A. Intelligence Explosion: Evidence and Import. In *Singularity Hypotheses*; Springer: Berlin/Heidelberg, Germany, 2012.
50. Christiano, P. Prosaic AI Alignment. Available online: https://ai-alignment.com/prosaic-ai-control-b959644d79c2 (accessed on 18 February 2019).
51. Itut Reality Check: 'We Are Not Nearly As Close To Strong AI As Many Believe'. Available online: https://news.itu.int/reality-check-not-nearly-close-strong-ai-many-believe/ (accessed on 18 February 2019).
52. Partnership for AI. Available online: https://www.partnershiponai.org/ (accessed on 18 February 2019).
53. Turchin, A. Human Upload as AI Nanny 2017. Available online: https://www.academia.edu/38386976/Human_upload_as_AI_Nanny (accessed on 19 February 2019).
54. Mahoney, M. A Proposed Design for Distributed Artificial General Intelligence. 2008. Available online: http://mattmahoney.net/agi2.html (accessed on 18 February 2019).
55. Alexander, S. Ascended Economy? Available online: http://slatestarcodex.com/2016/05/30/ascended-economy/ (accessed on 18 February 2019).
56. Hanson, R.; Sun, W. Probability and Asset Updating using Bayesian Networks for Combinatorial Prediction Markets. *arXiv* **2012**, arXiv:1210.4900.
57. Camarinha-Matos, L.M.; Afsarmanesh, H. Collaborative networks: a new scientific discipline. *J. Intell. Manuf.* **2005**, *16*, 439–452. [CrossRef]
58. Luksha, P. NeuroWeb Roadmap: Results of Foresight & Call for Action. 2014. Available online: https://dlib.si/details/URN:NBN:SI:DOC-IXKS9ZQW (accessed on 18 February 2019).
59. Vinding, M. *Reflections on Intelligence*; Heinemann: Portsmouth, NJ, USA, 2016.
60. Marx, K. Capital: A Critique of Political Economy. The Process of Production of Capital. 1867. Available online: https://oll.libertyfund.org/titles/marx-capital-a-critique-of-political-economy-volume-i-the-process-of-capitalist-production (accessed on 18 February 2019).
61. Smart, J. The transcension hypothesis: Sufficiently advanced civilizations invariably leave our universe, and implications for METI and SETI. *Acta Astronaut.* **2012**, *78*, 55–68. [CrossRef]

62. Gwern. Why Tool AIs want to be Agent AIs 2016. Available online: https://www.gwern.net/Tool-AI (accessed on 18 February 2019).
63. Yudkowsky, E. Harry Potter and Method of Rationality. 2010. Available online: https://fanlore.org/wiki/Harry_Potter_and_the_Methods_of_Rationality (accessed on 18 February 2019).
64. Bostrom, N.; Armstrong, S.; Shulman, C. Racing to the Precipice: a Model of Artificial Intelligence Development. *AI Soc.* **2013**, *31*, 201–206.
65. Shulman, C. Arms races and intelligence explosions. In *Singularity Hypotheses*; Springer: Berlin, Germany, 2011.
66. Bostrom, N. Strategic Implications of Openness in AI Development. *Glob. Policy* **2016**, *8*, 135–148. [CrossRef]
67. Baum, S.D. On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *Glob. Catastroph. Risk.* **2016**, *32*, 543–551. [CrossRef]
68. Ouagrham-Gormley, S.B. Dissuading Biological Weapons. In *Proliferation Pages*; Springer: Berlin, Germany, 2013; pp. 473–500.
69. Auerbach, D. The Most Terrifying Thought Experiment of All Time. Available online: http://www.slate.com/articles/technology/bitwise/2014/07/roko_s_basilisk_the_most_terrifying_thought_experiment_of_all_time.html (accessed on 18 February 2019).
70. Fernando, C. PathNet: Evolution Channels Gradient Descent in Super Neural Networks. *arXiv*, 2017; arXiv:1701.08734.
71. Nelson, R. How to Deter a Rogue AI by Using Your First-mover Advantage. 2007. Available online: http://www.sl4.org/archive/0708/16600.html (accessed on 18 February 2019).
72. Kharpal, A. *Elon Musk: Humans Must Merge with Machines or Become Irrelevant in AI Age*; CNBC: Englewood Cliffs, NJ, USA, 2017.
73. Brin, D. *The Transparent Society*; Perseus Book: New York, NY, USA, 1998.
74. Hanson, R. *The Age of Em: Work, Love, and Life when Robots Rule the Earth*; Oxford University Press: Oxford, UK, 2016.
75. Bostrom, N. *Hail Mary, Value Porosity, and Utility Diversification*; Oxford University Press: Oxford, UK, 2016.
76. Lem, S. The Investigation. 1959. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1651-2227.1959.tb05423.x (accessed on 18 February 2019).
77. Urban, T. Neuralink and the Brain's Magical Future. 2017. Available online: https://waitbutwhy.com/2017/04/neuralink.html (accessed on 18 February 2019).
78. Bostrom, N. Human genetic enhancements: a transhumanist perspective. *J. Value Inq.* **2003**, *37*, 493–506. [CrossRef]
79. Smith, P.D. *Doomsday Men: The Real Dr. Strangelove and the Dream of the Superweapon*; St. Martin's Press: New York, NY, USA, 2007.
80. Alexander, S. Should AI Be Open. Available online: https://slatestarcodex.com/2015/12/17/should-ai-be-open/ (accessed on 18 February 2019).
81. Baker, B.H. *The Gray Matter: The Forgotten Story of the Telephone*; Telepress: Kent, WA, USA, 2000; ISBN 0-615-11329-X.
82. The Telegraph Russian Spacecraft Landed on Moon Hours Before Americans. Available online: http://www.telegraph.co.uk:80/science/space/5737854/Russian-spacecraft-landed-on-moon-hours-before-Americans.html (accessed on 18 February 2019).
83. Venture Scanner Artificial Intelligence Q1 Update in 15 Visuals 2016. Available online: https://www.venturescanner.com/blog/2016/artificial-intelligence-q1-update-in-15-visuals (accessed on 18 February 2019).
84. Yampolskiy, R. From Seed AI to Technological Singularity via Recursively Self-Improving Software. *arXiv* **2015**, arXiv:1502.06512.
85. Drexler, E.; Phoenix, C. Safe exponential manufacturing. *Nanotechnology* **2004**, *15*, 869.
86. Bontchev, V. *Are Good Computer Viruses Still a Bad Idea?* EICAR: London, UK, 1994.
87. Sotala, K.; Valpola, H. Coalescing minds: brain uploading-related group mind scenarios. *Int. J. Mach. Conscious.* **2012**, *4*, 293–312. [CrossRef]
88. Batin, M.; Turchin, A.; Markov, S.; Zhila, A.; Denkenberger, D. Artificial Intelligence in Life Extension: From Deep Learning to Superintelligence. *Inform. Slov.* **2018**, *41*, 401.
89. Alexander, S. Book Review: Age of Em. Available online: http://slatestarcodex.com/2016/05/28/book-review-age-of-em/ (accessed on 18 February 2019).
90. Bostrom, N. Are You Living in a Computer Simulation? *Publ. Philos. Q.* **2003**, *53*, 243–255. [CrossRef]

91. Omohundro, S. The basic AI drives. In Proceedings of the AGI Frontiers in Artificial Intelligence and Applications, Memphis, TN, USA, 1–3 March 2008.
92. Bostrom, N. Existential risk prevention as global priority. *Glob. Policy* **2013**, *4*, 15–31. [CrossRef]
93. Shakirov, V. Review of State-of-the-Arts in Artificial Intelligence with Application to AI Safety Problem. *arXiv* **2016**, arXiv:1605.04232.
94. DeepMind AlphaGo. Available online: https://deepmind.com/research/alphago/ (accessed on 18 February 2019).
95. Ministry of National Defense of the People's Republic of China. *The Dawn of the Intelligent Military Revolution*; Ministry of National Defense of the People's Republic of China: Beijing, China, 2016.
96. Factored Cognition (May 2018) Ought. Available online: https://ought.org/presentations/factored-cognition-2018-05 (accessed on 25 January 2019).

big data and
cognitive computing

Article

Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach

Steven Umbrello

Institute for Ethics and Emerging Technologies, Via San Massimo 4, Turin 10123, Italy; Steve@ieet.org

Received: 18 December 2018; Accepted: 2 January 2019; Published: 6 January 2019

Abstract: This paper argues that the Value Sensitive Design (VSD) methodology provides a principled approach to embedding common values into AI systems both early and throughout the design process. To do so, it draws on an important case study: the evidence and final report of the UK Select Committee on Artificial Intelligence. This empirical investigation shows that the different and often disparate stakeholder groups that are implicated in AI design and use share some common values that can be used to further strengthen design coordination efforts. VSD is shown to be both able to distill these common values as well as provide a framework for stakeholder coordination.

Keywords: value sensitive design; VSD; design for values; safe for design; AI; ethics

1. Introduction

Value Sensitive Design (VSD) is a design methodology that begins with the premise that technologies are value-laden and that human values are continually implemented both during and after the design of a technology [1,2]. The 'sensitivity' of VSD is to the values that are held by the multitude of stakeholders that are both directly and indirectly enrolled during technological design whether they be engineers, CEOs and/or the relevant publics. This paper aims to argue for the VSD approach as a potentially suitable methodology for artificial intelligence coordination between the often-disparate publics, governmental bodies, and industry. In evaluating the applicability of VSD to AI coordination, this paper eschews any in-depth discussion of superintelligence or AI risk scenarios. In doing so, the aim of this paper is to lay out arguments for the adoption of VSD that can have an immediate impact on existing AI systems and on the systems of the near future. The value of this immediacy is taken for granted given the urgency proposed by the abundant AI risk research.

VSD exists among various other safe-by-design methodologies within the field of responsible research and innovation (RRI) and itself comes in various forms depending on the domain of applications [3–6]. It is largely agreed in the design literature, spanning back to the inception of technologies studies that technology is not value-neutral, but rather that values are consistently implicated in design [7,8]. Artificial intelligence, like robotics, nanotechnology, and information and communication technologies (ICTs), among others, is a sociotechnical structure that implicates not only the physical, or digital entity itself, but also the infrastructures, people and politics that it emerges from and into [9–14]. Not only this, but sociotechnical systems function only in accordance with the boundaries of this social context, they require actors and institutions that constrain and direct developmental pathways towards certain avenues rather than others [15,16]. The actors and infrastructures that allow a sociotechnical system to emerge naturally implicate values with questions such as: which funding bodies are permitted to distribute monies? How are research avenues chosen and who judges what is an acceptable research stream? How are opportunity-cost decisions made and under what criteria are some paths chosen rather than others? Because each of these questions is naturally implicated in design and because each of them implicates values, values in design must be considered more carefully, not only of the technologies in question themselves but also the institutions and social infrastructures that enroll these values.

VSD provides such a way to evaluate the values that are implicated both on technical and social dimensions as has been demonstrated in its application for other socio-technical systems [17–20]. Dignum et al. (2016) and Oosterlaken (2015) both explore the potential application of applying the VSD framework to socio-technical energy systems, whereas Umbrello and De Bellis (2018) explore more explicitly the potential boons that a VSD approach can bear on the technical development of intelligent agents (IA). Umbrello and De Bellis (2018) provide a theoretical basis for which moral values of stakeholders could be designed into the technical systems of IAs and provides means for adjudicating moral overload [21], however, they do not give any real account of how VSD could ameliorate the gap between various, often conflicting stakeholders. Dignum et al. (2016), however, provide a valuable analysis of various groups such as the federal government, non-governmental organizations (NGOs) and commercial organizations with regards to the surveying and extraction of shale gas in the Netherlands. In evaluating the policy documents of these different stakeholders, the authors were able to infer and distill a set of root values. However, although both Dignum et al. (2016) and Oosterlaken (2015) provide useful studies, they do not give any empirical case for the application of VSD to existing sociotechnical systems. Mouter, Geest, and Doorn (2018) argue that because the Dutch government scuttled the exploitation of the shale gas in the Netherlands, there was no way for Dignum et al. (2016) to elicit the explicit design considerations that can be used for a thorough VSD analysis [22].

To the best of my knowledge, this paper is the first to evaluate the merits of the VSD framework for AI coordination per se. Prior literature on VSD has focused on its methodology [8,23], its application to existent technologies [24,25], its philosophical underpinnings [26,27] and even to the reduction of future AI risk [20]. These studies provide useful information regarding both VSD and AI but do not provide any tangible analysis of the issues of coordination, nor to those that are particular to AI. This paper's application of the VSD approach as a means to ameliorate the often-disparate stakeholders that are implicated in the development and use of AI technologies is particularly unique. It is similarly the intent of this paper to spark further research on some of the issues regarding how VSD can be used to coordinate stakeholders of other technological innovations that converge with AI, such as nanotechnology and biotechnology.

To successfully tackle this argument, this article is organized into the following sections (see graphical abstract): the first section will lay out the methodological framework of the VSD approach as well as how it has been applied to other technological innovations. In doing so, one can begin to conceptualize both the strengths and potential drawbacks of the VSD approach as it can be formulated for application to AI systems. The second section will draw upon the work done in §1 by beginning to sketch multiple pathways for potential AI coordination by formulating specific examples of coordination between various AI stakeholders by drawing on a specific case study that implicates a variety of stakeholders. In doing so, this paper builds on the previous work done by Umbrello and De Bellis (2018) which explores how the VSD approach can be used to design intelligent agents (IAs) specifically. While that paper explored the technicalities of IA design, this paper investigates the stakeholders themselves to better form pathways for coordination. The final section of this paper sketches broader theoretical implications that these conclusions may have and points to potential future research avenues.

2. Material and Methods

Emerging from the domain of human-computer interaction (HCI) and ICT, VSD has since developed into a largely adopted design approach to incorporate human values (and perhaps even non-human) values during both the early and latter design phases of technologies [23,28]. Since its inceptions in the early 1990s, VSD has been adopted as a proposed framework for the design of identity technologies [25], energy technologies such as wind turbines [19,24], robotics and autonomous agents such as care robots, autonomous vehicles, and AI in the medical field [20,29–32], information and communication technologies such as sensors and communicative computer software [33–38], health technologies such as ambulatory therapeutic assistance systems and seizure detectors [39–42], and nanotechnology both in its advanced and contemporary forms [43–45]. VSD is described by its founders Batya Friedman et al., as a tripartite framework consisting of conceptual, empirical and technical investigations [23].

Conceptual investigations are characterized as philosophical evaluations of determining who the stakeholders are, determining the values that are identified, what values should be chosen, as well as how conflicts between values are to be resolved. Next, empirical investigations use various surveying methods such as observations and interviews, as well as other explorative tools to determine if the values distilled in conceptual investigations can be successfully embedded into a certain technological design [1]. The third investigation, technical investigations, is characterized by two steps: the first determines how the technology under question constrains or supports humans values whereas the second avenue determines how the distilled values of the conceptual investigations can be sufficiently embedded in the technological design [46]. Although empirical and technical investigations are complimentary and akin to one another, the difference between the two is not insignificant. Empirical investigations focus primarily on stakeholders who are affected, either directly or indirectly by the technological design whereas technical investigations investigate the technology per se.

VSD is often chosen over competing theories because its emphasis is not only on the conceptualization of the values that are embedded, or aim to be embedded in a design, but because it requires adding an empirical and technical analysis to evaluate the role of systems and institutions that affect design as well as how stakeholder groups form a co-constitutive role in a technologies safe-adoption [47]. The importance here for AI stakeholders is that VSD provides a principled way of engaging with different stakeholder groups, giving a way for their values and perceptions of AI to be formulated into a root set of instrumental values that can then be brought directly into the design process. Lastly, the framework may tally benefits to the design practice by determining moral overload a priori, establishing understanding within and between stakeholder groups regarding potentially emerging value-conflicts. Moral overload in the design literature refers to when elicited stakeholders provide conflicting, yet still important values for technological design [21]. What VSD does not do however is provide a clear way of *actually* embedding values into a design. Its aim is to highlight the root values at play by stakeholders and to determine if the technology in question supports or constrains those values [48], however, formulated the notion of a 'value hierarchy' (see Figure 1) that allows the moral values of stakeholders to be more easily conceptualized as functional design requirements [48].

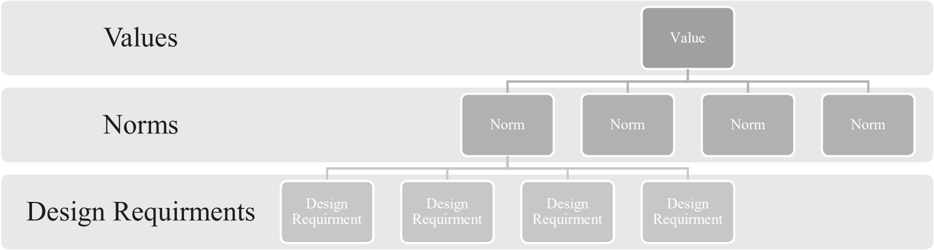


Figure 1. Top-Down Values hierarchy (Source: [48]).

What this paper does then, in order to better conceptualize how different stakeholders relevant to AI conceptualize values, is use Van de Poel's value hierarchy as the main tool to construct a set of root values that can aid to bridge the cooperative design gap. A top-down hierarchy of values such as Figure 1 consists of three distinct ranks, the top rank (Values) is objective. It is objective in the sense that the root values distilled are not sensitive to context [27] or culture. For example, [26] argues against this very notion, arguing for both intersubjectivity as a means by which to reconceptualize

VSD as well as the reformulation of VSD away from moral law theories towards an imaginative theory of morality that is more in line with modern neuroscience. The proceeding rank consists of norms, which inhere as every form of imperative or constraint on action, these differ from the root values of the higher-order rank of values because norms are sensitive to context and situation. The lowest rank aims to formalize the higher-order rank of norms as functional design requirements. In doing so, the norms aim to be translated into an applied practice that can then be introduced into the design flow [48–50].

However, the hierarchy need not flow in the top-down direction as the original formulators of VSD originally conceptualized; it can similarly move from the bottom upwards. It begins naturally with a particular set of existing design requirements that are then used to distill a common set of root values. The following section of this paper employs this dual-directional analysis (best conceptualized by Figure 2) to better find a path of cooperation between AI stakeholders.

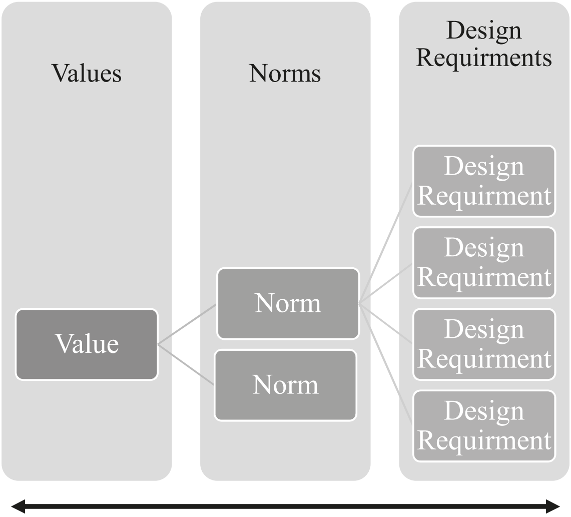


Figure 2. Bi-Directional Values Hierarchy.

The purpose of this paper is to determine the suitability of the VSD approach to the coordination of various stakeholders involved and implicated in beneficial AI [51] research and development. In doing so, it draws upon one potentially controversial case, that of the appointment of the UK Select Committee on Artificial Intelligence. This particular case has been selected over other controversial cases because (1) its ad hoc nature gives it a discrete time-specificity and ease by which the case can be analyzed, and (2) the case did and continues to garner media scrutiny. Because of both (1) and (2), coupled with the potential societal influence that the committee can have as a result; the ability to source relevant material and literature is straightforward and accessible.

In the second report of the 2016–17 session of the House of Lords Liaison Committee—an advisory group to the House which advises, oversees, and reviews the resources needed for the selection and coordination of select committees and ad hoc committees—advised for the formation of four ad hoc committees of which the subject of one was solely to focus on artificial intelligence [52]. These ad hoc committees, selected in the 2016–17 session, were established as year-long seats, which were then to report their findings in time for the 2017–18 session in March 2018.

Acknowledging the impacts of continued technological advances, proposals for the establishment of an ad hoc select committee on artificial intelligence were forwarded to focus on the economic, social and ethical issues implicated by the design and use of artificial intelligence systems. Because it is a topic of specific interest that does not fall within the purview of the expertise of any existing committee (i.e., it is the first of its kind), the establishment of a topic-specific committee was decided upon. More specifically, the ad hoc committee was envisioned to evaluate the impact of AI on the following topics, taking into account both the arguments of the 'techno-optimists' and the 'techno-pessimists':

- Pace of technological change
    - ○ Relationship between developments in artificial intelligence and productivity growth;
    - ○ Creation of new jobs;
    - ○ Sectors and occupations most at threat from automation.
- Economic and social issues
    - ○ The role of government in the event of widespread job displacement;
    - ○ Further education and training, for both children and adults;
    - ○ Unemployment support, including the case for a universal basic income;
    - ○ Government funding for artificial intelligence-related research and development.
- Ethical issues
    - ○ The government's role in monitoring the safety and fairness of artificial intelligence;
    - ○ Transparency around the use of 'big data';
    - ○ Privacy rights of individuals;
    - ○ General principles for the development and application of artificial intelligence. (Source: [52])

From 29 June 2017, when the appointments of the Select Committee on AI were established, the members met in three closed sessions over the course of the month. The following meeting was their visit to DeepMind on 13 September 2017. The following months consisted of a combination of both closed private sessions as well as public evidence sessions of which transcripts of the panels are fully accessible online [53]. After several closed sessions between January and March 2018, the Select Committee's final report was published on 16 April 2018 and later publicly debated in government on 19 November 2018.

The final report concluded that the UK is well positioned to be a global leader in AI research and development. Properly designed and implemented, the report considered the UK to be in a unique position to address social, economic and ethical issues that existed and that may arise with the design and implementation of AI system and take advantage of the economic and social benefits that they are predicted to usher. Similarly, the report acknowledges the value-ladenness of technologies, their socialtechnicity, and the past issues of prejudice being designed into technological systems; the resolution was taking care in the early design phases to ensure an equitable design process.

Finally, the report argues for more transparent access to data and the enrollment of stakeholders into the decision-making processes of industry and governmental bodies directly responsible for the design of AI. Presently, discussions of practical steps to bridge cooperative gaps are taking place to apply the recommendations of the committee's report.

As already outlined, the VSD approach was originally construed as an anticipatory design framework that envisioned a technological design in isolation from the socialtechnicity that it was to emerge in. However, the already widespread use of AI systems makes a purely ex-ante approach impotent, and for this reason, both the top-down and bottom-up rankings are required. These permit adjustments and modifications as new information makes itself known [54].

To this end, in this section, I uncover some of the most pertinent values of ethical importance within the context of this case. Typically, as per the original instantiations of the VSD approach, the vast body of philosophical and sociological literature is levied to better distill a set of core values. Friedman et al., along with [20] provide a strong point of departure within the realm of both HCI and AI regarding potentially relevant values such as safety, privacy, accountability, and sustainability [20,55]. The remainder of the list of values (Table 1) is drawn from the various written and oral transcripts that eventually formed the collated evidence volumes that were gathered by the Select Committee [56]. As such, what follows is an *empirical investigation* as per the VSD approach given by the committee themselves engaged in the conceptual investigations of determining the ethical values implicated in AI.

The written comprehensive evidence volume consists of 223 separate reports by policy experts, academics, NGOs, think tanks, governmental bodies, and industry leaders [56]. This categorization employed in this paper to separate the different evidence reports and testimonies is taken directly from the reports themselves which are explicit in their affiliation and category. Similarly, the oral evidence volume consists of 57 separate oral testimonies by similar groups and individuals [57]. Likewise, the government response to the House of Lords Artificial Intelligence Select Committee's report provides a clear perspective on how the UK aims to address the report's findings [58]. What should be noted here is that the represented sample size garnered by the reports (and by the committee's search) do not reflect a full sample size of stakeholders affected (or can be affected indirectly) by AI technologies. The values distilled are those projected by the 'experts' appointed by the committee to draw reports. Because of this, this paper, as well as the case study as a whole, represent an initial sketch of how conceptual investigations can be undertaken, and are an illustration of the further work that needs to be done in order to draw a representative stakeholder group that accounts for population from the considered area, in what concerns its structure: age, gender, occupation, educational level, family size.

The bi-directional approach to distilling values and design requirements is of particular use when investigating these documents given that their eclectic sources, ranging from not only those listed but also those with both philosophical and engineering backgrounds. The ability to use both approaches to come to a similar set of values and design requirements permits a more thorough approach to determining's a common list of values, even if it only serves as a starting point for collaborative actions between the relevant stakeholders implicated in the government's proceedings.

3. Results

To this end, the list of values in Table 1 is the result of a prolonged distillation of the bi-directional method. Each of the 223 separate written evidence reports, as well as the transcripts of the 57 oral witness testimonies, were read for both an explicit account of what needed to be construed as a design requirement (i.e., a value) whereas norms and technical design requirements were contextualized into values. What resulted is a major overlap of a series of 12 values ranging in support. Transparency was shown to be the most widely supported, overlapping with 146 different reports. The majority of the evidence reports employed the term transparency, while others preferred interpretability or 'explainability', sometimes interchangeably. The final report opted for the use of 'intelligibility' to refer to the broader issue. Similarly, intelligibility can be approached in two distinct ways: (1) technical transparency and (2) Explainability. Similarly, control and data privacy came in both second and third, respectively, in terms of support by the different evidence reports (see Figure 3 for the rank-order distribution).

Prescriptions for technical *transparency* to permit users and designers to understand how and why the decisions made by AI systems were taken was one of the most identified top-down values. Technical recommendations, like the ability for both users and designers to access a system's source code, were the primary norms identified, however, that, per se, does not entail transparency for why certain decisions were chosen over others, nor does it show the data input that leads to those decisions. Similarly, transparency was argued to be a value that is contingent on the stakeholder group in question, as well as the purpose of the AI system in question. For example, Professor Chris Reed, Professor of Electronic Commerce Law, Queen Mary University of London, argued that:

There is an important distinction to be made between ex-ante transparency, where the decision-making process can be explained in advance of the AI being used, and ex-post transparency, where the decision-making process is not known in advance but can be discovered by testing the AI's performance in the same circumstances. Any law mandating transparency needs to make it clear which kind of transparency is required [59].

Table 1. The 12 values supported throughout the collected evidence volumes with the number of unique reports that explicitly supported each value or provided a design requirement or norm that could be distilled into a value.

Values	Academics/ Scholars/Universities	NGOs/Think Tanks/Non-Profits	Governmental Bodies	Industry/For Profit
Data Privacy	9	5	5	14
Accessibility	4	5	3	7
Responsibility	41	15	4	18
Accountability	35	10	10	25
Transparency	62	27	13	44
Explainability	5	4	2	4
Efficiency	19	16	5	19
Consent	35	25	5	19
Inclusivity	7	4	5	7
Diversity	26	14	5	24
Security	44	33	9	35
Control	65	35	7	34

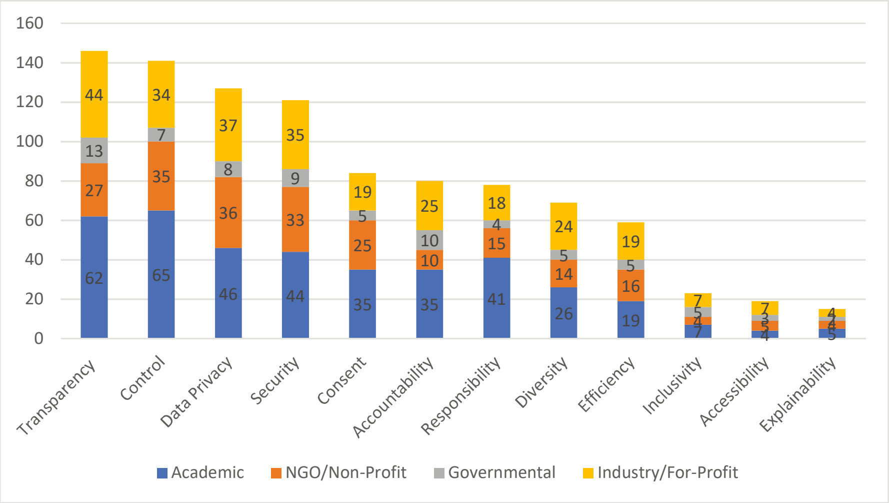


Figure 3. Rank-Order Distribution of Values. Numerical values represent individual report.

Certain constraints on ex-ante transparency thus could be warranted because absolute transparency prior to an AI development could severely curtain AI development and innovations. Nonetheless, sacrifice to innovation in favor of transparency was universally affirmed by the reports where fundamental human rights were at stake.

Diversity and inclusivity, on the other hand, were values that were identified through the bottom-up approach, usually in relation to a more explicit value and how that value can be strengthened or realized through design requirements. The value of transparency, for example, can help to determine what inputs are being fed into a system and determine if those inputs and the subsequent decisions are impartial, inclusive and diverse. These two values, in particular, were not identified in the top-down approach and were relegated exclusive to design requirements that supported more explicit norms and values.

4. Discussion

So far, this paper has looked at how a specific case study has engaged in conceptual investigations on AI design and development to determine the human values that are important to different stakeholders. Values were identified both from the top-down and bottom-up methods. What follows in this section is a cursory look at how VSD can be further harmonized with the existent and ongoing work in AI to further bridge disparate stakeholder groups.

Transparency, control, and privacy arose in this study as the most explicit values expressed, while values such as diversity, inclusivity, and accessibility were expressed as bottom-up design requirements or norms that were related to securing one of those three values. Because of this, those values, particularly transparency, is used to discuss how the VSD approach could be used to further embed that value into AI design.

In evaluating the content that discussed transparency either explicitly or as a design requirement, the concerns that were mostly expressed were that ex-post technical-approaches to AI systems' transparency is difficult, if not impossible. However, there are nonetheless cases where such transparency is imperative, even if they come at the cost of "power and accuracy" [59]. To this end, transparency can be affirmed through the design requirement of technical explainability, in which ex-ante approaches to systems development require AIs to continually explain the logic and inputs used to arrive at their decisions [60]. The adoption of the VSD approach during preliminary stages of AI development thus can help to mitigate the difficulties of ex-post black boxes and help to determine the level of stakeholder tolerance between competing values such as transparency and privacy. For this reason, the inclusion of foundation norms such as "determining the diversity and inclusivity of data sets" helps to strengthen higher rank-ordered values such as transparency. The inclusion of these norms throughout the design process provides both a path for the formalization of new design requirements, as well as a way to reformulate values in less-obvious ways.

Additionally, the values distilled in both this study, as well as in the collated report should not discount, nor be prioritized over those of continued conceptual investigations by designers. The investigations of values as a purely conceptual, a priori practice aids designers to deliberate on values that may not emerge in stakeholder elicitations. Although the design of AI systems with the explicit values of stakeholders may increase system adoption and acceptance, the values that can emerge through the principled conceptual investigations that VSD formalizes is also of importance. Similarly, given the socio-technicity of AI, stakeholders may often overlook how infrastructures, technical standards, the values of designers, and other social systems constitute and shape the values that are implicated in technological development. Similarly, delimiting who the stakeholders are and adequately selecting a representative group to elicit values is difficult, hence making conceptual investigations an important step along with empirical and technical investigations. In doing so, when designers elicit stakeholder involvement, they can then reflect on the values of conceptual investigations to continually adapt them to the changing technical and empirical input.

Although VSD does not offer the ideal solution for bridging stakeholder groups and solidifying their coordination in the design of AI, it does nonetheless present the fundamentals for (1) determining common values across stakeholder groups through both norms and design requirements (and vice versa) and (2) makes value conflicts functionally apparent and addressable thus (3) permitting both ex ante and ex post interventions to take place that account for a wide variety of stakeholder values. Having a formalized approach like this, with clear stages and delineations, allows designers to design AI systems in a principled way that reduces the likelihood of biased or uninformed decisions. A step that can bet taken by committees and similar groups such as the UK Select Committee on AI is to acknowledge a common set of values amongst the select stakeholders, extend those conceptual and empirical investigations to other stakeholder groups that were perhaps not considered during the initial conceptual investigations and determine if there is any overlap. Similarly, those values can then be used to determine design requirements that can express those values at technical level in design.

5. Conclusions

The purpose of this paper was to explore the potential applicability of the VSD methodology to the development and fostering of cooperation and collaboration between various stakeholder communities in the design and development of AI systems. Through the application of empirical investigations as outlined in the VSD framework, this paper explored the implicated human values that may be relevant to the design of AI systems. It concluded that, in the case of the UK Select Committee on AI, that a common value hierarchy could be distilled from disparate stakeholder groups and from different mediums of translation (i.e., reports, testimonies, and newspapers). The bi-directional approach to the value-hierarchy was shown to be the best way to distill both values and design requirements given that different mediums offered different ways of arriving at either one (policy reports vs. news reports). Transparency, for example, was always shown through the top-down approach whereas values such as diversity and inclusivity were only through the bottom-up approach. An important observation of this study is that transparency is an important, yet multi-faceted and often difficult, value to incorporate into design, requiring ex-ante interventions at the design stages to increase transparency via technical explainability.

The findings of this paper have the potential to allow both stakeholders and engineers to better conceptualize the values of different groups that may reduce AI recalcitrance and increase stakeholder *inclusivity* and *accessibility*. In doing so, the design process for the multitude of AI systems can be strengthened both from the early design phases and throughout their development through continued stakeholder dialogue.

It is acknowledged that both this paper and VSD have their limitations. The investigations carried out in this particular case study are both socially and culturally situated, and thus limited. Similarly, the values explored by VSD are considered universal rather than socially or culturally relative [26]. Likewise, VSD affirms strong anthropocentrism in its value investigations whereas an abundance of literature from both cultural anthropology and philosophical ecology have shown that the values of nonhuman actors (and perhaps eventually AGI/ASI) are always already implicated in human actions in the Anthropocene [61,62]. This study has shown from where initial steps can be taken towards the design of beneficial AI, but further research studies should not only work from the initial premises of this paper but explore the viability of both non-anthropocentric values as well as the flexibility of the underlying assumptions of VSD's conceptual investigations. Although some recent work has begun these investigations [26–28], it has yet to be adopted as common practice within the design scholarship and requires further argumentation if it is to be so.

Additionally, VSD can be limited in many cases by constraints on the relevant literature to undertake conceptual investigations. Similarly, restricted access to relevant stakeholder groups, diversity, and inclusivity of the members of those groups and the ability to resolve the moral overload of value conflicts in a clear and principled way all limit the VSD methodology. This paper, for example, is not only limited in these ways but it also focuses primarily on empirical investigations and disregards the technical investigations that are critical to VSD.

Nonetheless, what this study has shown is that VSD can be applied both ex-ant and ex-post facto to sociotechnical systems that already exist. What is needed are both research and policy measures that can determine the actual impact of the adoption of VSD as a general framework for design. What VSD aims to do, and this paper should have shown more explicitly, is that through a thorough investigation of various sources and stakeholders, various design requirements can be translated into a common set of held values and that explicit values can also be translated into design requirements. Similarly, the work that has gone into this study to better facilitate the hierarchy of values from these various mediums shows that that VSD methodology with a bi-directional hierarchy approach requires a substantial time investment to ensure that important values or design requirements are not passed over. Whether this is true for various cultures and social contexts is yet to be seen and can only be done with its wider adoption, if and when that happens. That being said, continued VSD research should similarly look at the situations in which the produced studies emerge to better determine weakness within both the studies themselves and the VSD framework (i.e., improvements could reduce partiality and cultural bias, and give voice to silenced stakeholders).

Funding: This research received no external funding.

Acknowledgments: I would like to thank the two anonymous peer reviewers who have taken the time to both quickly and thoroughly review this manuscript. Their recommendations have undoubtedly increased the quality of this paper. Any remaining errors are the author's alone. The views in the paper are the authors' alone and not the views of the Institute for Ethics and Emerging Technologies.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Friedman, B.; Kahn, P.H.; Borning, A.; Huldtgren, A. Value Sensitive Design and Information Systems. In *Early Engagement and New Technologies: Opening up the Laboratory*; Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 55–95.
2. Van den Hoven, J. The Design Turn in Applied Ethics. In *Designing in Ethics*; van den Hoven, J., Miller, S., Pogge, T., Eds.; Cambridge University Press: Cambridge, UK, 2017; pp. 11–31.
3. Boenink, M. The Multiple Practices of Doing 'Ethics in the Laboratory': A Mid-Level Perspective. In *Ethics on the Laboratory Floor*; Springer: Berlin, Germany, 2013; pp. 57–78.
4. Doorn, N.; Schuurbiers, D.; Van de Poel, I.; Gorman, M.E. *Early Engagement and New Technologies: Opening up the Laboratory*; Springer: Berlin, Germany, 2014; Volume 16.
5. Fisher, E.; O'Rourke, M.; Evans, R.; Kennedy, E.B.; Gorman, M.E.; Seager, T.P. Mapping the Integrative Field: Taking Stock of Socio-Technical Collaborations. *J. Responsib. Innov.* **2015**, *2*, 39–61. [CrossRef]
6. Micheletti, C.; Benetti, F. Safe-by-Design Nanotechnology for Safer Cultural Heritage Restoration. Available online: http://atlasofscience.org/safe-by-design-nanotechnology-for-safer-cultural-heritage-restoration/ (accessed on 15 December 2017).
7. Winner, L. Do Artifacts Have Politics? *Technol. Future* **2003**, *109*, 148–164. [CrossRef]
8. Van den Hoven, J.; Manders-Huits, N. Value-Sensitive Design. In *A Companion to the Philosophy of Technology*; Wiley-Blackwell: Oxford, UK, 2009; pp. 477–480.
9. Bijker, W.E.; Hughes, T.P.; Pinch, T. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*; MIT Press: Cambridge, MA, USA, 1987.
10. Bechtold, U.; Fuchs, D.; Gudowsky, N. Imagining Socio-Technical Futures—Challenges and Opportunities for Technology Assessment. *J. Responsib. Innov.* **2017**, *9460*, 1–15. [CrossRef]
11. Pitt, J.; Diaconescu, A. Interactive Self-Governance and Value-Sensitive Design for Self-Organising Socio-Technical Systems. In Proceedings of the 2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS*W), Augsburg, Germany, 12–16 September 2016; pp. 30–35. [CrossRef]
12. Baxter, G.; Sommerville, I. Socio-Technical Systems: From Design Methods to Systems Engineering. *Interact. Comput.* **2011**, *23*, 4–17. [CrossRef]
13. Trist, E. The Evolution of Socio-Technical Systems. *Occas. Pap.* **1981**, *2*, 1981.
14. Crabu, S. Nanomedicine in the Making. Expectations, Scientific Narrations and Materiality. *TECNOSCIENZA Ital. J. Sci. Technol. Stud.* **2014**, *5*, 43–66.
15. Williamson, O.E. Transaction-Cost Economics: The Governance of Contractual Relations. *J. Law Econ.* **1979**, *22*, 233–261. [CrossRef]
16. Wüstenhagen, R.; Wolsink, M.; Bürer, M.J. Social Acceptance of Renewable Energy Innovation: An Introduction to the Concept. *Energy Policy* **2007**, *35*, 2683–2691. [CrossRef]
17. Künneke, R.; Mehos, D.C.; Hillerbrand, R.; Hemmes, K. Understanding Values Embedded in Offshore Wind Energy Systems: Toward a Purposeful Institutional and Technological Design. *Environ. Sci. Policy* **2015**, *53*, 118–129. [CrossRef]
18. Dignum, M.; Correljé, A.; Cuppen, E.; Pesch, U.; Taebi, B. Contested Technologies and Design for Values: The Case of Shale Gas. *Sci. Eng. Ethics* **2016**, *22*, 1171–1191. [CrossRef]

19. Oosterlaken, I. Applying Value Sensitive Design (VSD) to Wind Turbines and Wind Parks: An Exploration. *Sci. Eng. Ethics* **2014**, *21*, 359–379. [CrossRef] [PubMed]
20. Umbrello, S.; De Bellis, A.F. A Value-Sensitive Design Approach to Intelligent Agents. In *Artificial Intelligence Safety and Security*; Yampolskiy, R.V., Ed.; CRC Press: Boca Raton, FL, USA, 2018; pp. 395–410.
21. Van den Hoven, J.; Lokhorst, G.J.; van de Poel, I. Engineering and the Problem of Moral Overload. *Sci. Eng. Ethics* **2012**, *18*, 143–155. [CrossRef] [PubMed]
22. Mouter, N.; de Geest, A.; Doorn, N. A Values-Based Approach to Energy Controversies: Value-Sensitive Design Applied to the Groningen Gas Controversy in the Netherlands. *Energy Policy* **2018**, *122*, 639–648. [CrossRef]
23. Friedman, B.; Hendry, D.G.; Borning, A. A Survey of Value Sensitive Design Methods. *Found. Trends®Hum.–Comput. Interact.* **2017**, *11*, 63–125. [CrossRef]
24. Correljé, A.; Cuppen, E.; Dignum, M.; Pesch, U.; Taebi, B. Responsible Innovation in Energy Projects: Values in the Design of Technologies, Institutions and Stakeholder Interactions 1 (Draft Version for Forthcoming Book) Aad Correljé, Eefje Cuppen, Marloes Dignum, Udo Pesch & Behnam Taebi. In *Responsible Innovation 2*; Koops, B.-J., Oosterlaken, I., Romijn, H., Swierstra, T., van den Hoven, J., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 183–200.
25. Briggs, P.; Thomas, L. An Inclusive, Value Sensitive Design Perspective on Future Identity Technologies. *ACM Trans. Comput. Interact.* **2015**, *22*, 1–28. [CrossRef]
26. Umbrello, S. The Moral Psychology of Value Sensitive Design: The Methodological Issues of Moral Intuitions for Responsible Innovation. *J. Responsib. Innov.* **2018**, *5*, 186–200. [CrossRef]
27. Umbrello, S. Imaginative Value Sensitive Design: How Moral Imagination Exceeds Moral Law Theories in Informing Responsible Innovation. Masters Thesis, University of Edinburgh, Edinburgh, UK, 2018. [CrossRef]
28. Umbrello, S. Safe-(for Whom?)-By-Design: Adopting a Posthumanist Ethics for Technology Design. Masters Thesis, York University, Toronto, ON, Canada, 2018. [CrossRef]
29. van Wynsberghe, A. Designing Robots for Care: Care Centered Value-Sensitive Design. *Sci. Eng. Ethics* **2013**, *19*, 407–433. [CrossRef] [PubMed]
30. Santoni de Sio, F.; van den Hoven, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot. AI* **2018**, *5*, 15. [CrossRef]
31. Thornton, S.M.; Lewis, F.E.; Zhang, V.; Kochenderfer, M.J.; Gerdes, J.C. Value Sensitive Design for Autonomous Vehicle Motion Planning. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 1157–1162.
32. Dadgar, M.; Joshi, K.D. The Role of Information and Communication Technology in Self-Management of Chronic Diseases: An Empirical Investigation through Value Sensitive Design. *J. Assoc. Inf. Syst.* **2018**, *19*, 86–112. [CrossRef]
33. Dechesne, F.; Warnier, M.; van den Hoven, J. Ethical Requirements for Reconfigurable Sensor Technology: A Challenge for Value Sensitive Design. *Ethics Inf. Technol.* **2013**, *15*, 173–181. [CrossRef]
34. Warnier, M.; Dechesne, F.; Brazier, F. Design for the Value of Privacy. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–14.
35. Friedman, B. *Human Values and the Design of Computer Technology*; Friedman, B., Ed.; CSLI Publications: Stanford, CA, USA, 1997.
36. Huldtgren, A. Design for Values in ICT. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–24.
37. Van den Hoven, J. ICT and Value Sensitive Design. In *The Information Society: Innovation, Legitimacy, Ethics and Democracy In honor of Professor Jacques Berleur s.j.: Proceedings of the Conference "Information Society: Governance, Ethics and Social Consequences", University of Namur, Namur, Belgium, 22–23 May 20*; Goujon, P., Lavelle, S., Duquenoy, P., Kimppa, K., Laurent, V., Eds.; Springer: Boston, MA, USA, 2007; pp. 67–72. [CrossRef]
38. Weibert, A.; Randall, D.; Wulf, V. Extending Value Sensitive Design to Off-the-Shelf Technology: Lessons Learned from a Local Intercultural Computer Club. *Interact. Comput.* **2017**, *29*, 715–736. [CrossRef]

39. Mueller, M.; Heger, O.; Niehaves, B. Exploring Ethical Design Dimensions of a Physiotherapeutic MHealth Solution through Value Sensitive Design. In Proceedings of the Hawaii International Conference on System Sciences (HICSS), Maui, HI, USA, 16 August 2018.
40. Mueller, M.; Heger, O. Health at Any Cost? Investigating Ethical Dimensions and Potential Conflicts of an Ambulatory Therapeutic Assistance System through Value Sensitive Design. In Proceedings of the Thirty Ninth International Conference on Information Systems, San Francisco, CA, USA, 13–16 December 2018.
41. Van Andel, J.; Leijten, F.; Van Delden, H.; van Thiel, G. What Makes a Good Home-Based Nocturnal Seizure Detector? A Value Sensitive Design. *PLoS ONE* **2015**, *10*, e0121446. [CrossRef] [PubMed]
42. Cheon, E.; Su, N.M. Integrating Roboticist Values into a Value Sensitive Design Framework for Humanoid Robots. In Proceedings of the Eleventh ACM/IEEE International Conference on Human Robot Interaction, Christchurch, New Zealand, 7–10 March 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 375–382.
43. Timmermans, J.; Zhao, Y.; van den Hoven, J. Ethics and Nanopharmacy: Value Sensitive Design of New Drugs. *Nanoethics* **2011**, *5*, 269–283. [CrossRef] [PubMed]
44. van den Hoven, J. Nanotechnology and Privacy: The Instructive Case of RFID. In *Ethics and Emerging Technologies*; Sandler, R.L., Ed.; Palgrave Macmillan: London, UK, 2014; pp. 285–299.
45. Atomically Precise Manufacturing and Responsible Innovation: A Value Sensitive Design Approach to Explorative Nanophilosophy. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3141478 (accessed on 02 January 2019).
46. Friedman, B.; Howe, D.C.; Felten, E. Informed Consent in the Mozilla Browser: Implementing Value-Sensitive Design. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences, Big Island, HI, USA, 10 January 2002.
47. Doorn, N. Governance Experiments in Water Management: From Interests to Building Blocks. *Sci. Eng. Ethics* **2016**, *22*, 755–774. [CrossRef] [PubMed]
48. Van de Poel, I. *Translating Values into Design Requirements BT—Philosophy and Engineering: Reflections on Practice, Principles and Process*; Michelfelder, D.P., McCarthy, N., Goldberg, D.E., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 253–266.
49. Richardson, H.S. *Practical Reasoning about Final Ends*; Cambridge University Press: Cambridge, UK, 1997.
50. Vermaas, P.E.; Hekkert, P.; Manders-Huits, N.; Tromp, N. Design Methods in Design for Values. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; van den Hoven, J., Vermaas, P.E., van de Poel, I., Eds.; Springer: Dordrecht, The Netherlands, 2014; pp. 1–19.
51. Baum, S.D. On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *AI Soc.* **2017**, *32*, 543–551. [CrossRef]
52. Liaison Committee. *New Investigative Committees in the 2017–18 Session*; Authority of the House of Lords, The Stationery Office Limited: London, UK, 2017.
53. Lord Select Committee. Select Committee on Artificial Intelligence—Timeline. UK Parliament. Available online: https://www.parliament.uk/business/committees/committees-a-z/lords-select/ai-committee/timeline/ (accessed on 15 December 2018).
54. Franssen, M. Design for Values and Operator Roles in Sociotechnical Systems. In *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 117–149.
55. Friedman, B.; Kahn, P.H., Jr. Human Values, Ethics, and Design. In *The Human-Computer Interaction Handbook*; CRC Press: Boca Raton, FL, USA, 2007; pp. 1223–1248.
56. Lord Select Committee. *Select Committee on Artificial Intelligence. Collected Written Evidence Volume*; Lord Select Committee: London, UK, 2018.
57. Lord Select Committee. *Select Committee on Artificial Intelligence Collated Oral Evidence Volume*; Lord Select Committee: London, UK, 2018.
58. Secretary of State for Business, Energy and Industrial Strategy. *Government Response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?* Secretary of State for Business, Energy and Industrial Strategy: London, UK, 2018.
59. Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing and Able?* Select Committee on Artificial Intelligence: London, UK, 2018.

60. Hoff, R.D. Google Tries to Make Machine Learning a Little More Human. MIT Technology Review. Available online: https://www.technologyreview.com/s/542986/google-tries-to-make-machine-learning-a-little-more-human/ (accessed on 18 December 2018).
61. Harman, G. *Object-Oriented Ontology: A New Theory of Everything*; Penguin Random House: New York, NY, USA, 2018.
62. Morton, T. *Being Ecological*; MIT Press: Boston, MA, USA, 2018.

big data and cognitive computing

Article

Towards AI Welfare Science and Policies

Soenke Ziesche [1] and Roman Yampolskiy [2,*]

[1] Faculty of Engineering, Science and Technology, Maldives National University, Male' 20067, Maldives; soenke.ziesche@mnu.edu.mv

[2] Computer Engineering and Computer Science Department, University of Louisville, Louisville, KY 40292, USA

* Correspondence: roman.yampolskiy@louisville.edu; Tel.: +1-960-789-9304

Received: 24 November 2018; Accepted: 21 December 2018; Published: 27 December 2018

Abstract: In light of fast progress in the field of AI there is an urgent demand for AI policies. Bostrom et al. provide "a set of policy desiderata", out of which this article attempts to contribute to the "interests of digital minds". The focus is on two interests of potentially sentient digital minds: to avoid suffering and to have the freedom of choice about their deletion. Various challenges are considered, including the vast range of potential features of digital minds, the difficulties in assessing the interests and wellbeing of sentient digital minds, and the skepticism that such research may encounter. Prolegomena to abolish suffering of sentient digital minds as well as to measure and specify wellbeing of sentient digital minds are outlined by means of the new field of AI welfare science, which is derived from animal welfare science. The establishment of AI welfare science serves as a prerequisite for the formulation of AI welfare policies, which regulate the wellbeing of sentient digital minds. This article aims to contribute to sentiocentrism through inclusion, thus to policies for antispeciesism, as well as to AI safety, for which wellbeing of AIs would be a cornerstone.

Keywords: AI welfare science; AI welfare policies; sentiocentrism; antispeciesism; AI safety

1. Introduction

The purpose of this article is to contribute to the specification of policies towards the "interests of digital minds" within "a set of policy desiderata" outlined by Bostrom et al. [1] and further motivated by Dafoe [2].

A being is considered to have moral or intrinsic value, if the being is sentient, thus a moral patient. A being is sentient if it has the capacity to perceive qualia, including unpleasant qualia such as pain, which causes the being to suffer (humans and potentially other minds may also suffer for other reasons than unpleasant qualia, which is beyond the scope of this article). It is usually in the interest of sentient beings to avoid suffering. In addition to humans, many animals are considered to be sentient, which used to be controversial in the past, e.g., [3].

In this article, the focus is on sentient digital beings, mostly in the form of AIs, but sentient digital beings could also constitute subroutines [4], characters in video games or simulations [4–6], uploads of human minds [7]—e.g., through whole brain emulations [8]—or completely different sentient digital minds, as a subset of the vast overall space of minds [9]. While this topic is speculative and lacking evidence at this stage, the authors above and others argue that already now or in the future sentient digital beings or minds may exist, also e.g., [10–14]. An example for an opponent who does not believe in sentient digital beings is Dennett [15].

Furthermore, our premise is that digital beings may not only be sentient, but may also suffer (see also below a scenario for digital minds, which have exclusively pleasant perceptions and for which this article is largely not relevant). The suffering of any sentient being is a significant issue and may even increase in the future dramatically, which would also affect digital sentient beings [4] and to which a future superintelligence may contribute [16]. Therefore, it has been argued that the reduction of risks of future suffering of sentient beings deserves a higher priority [17].

This is interpreted as a non-zero probability for the existence of at least temporarily suffering sentient digital beings, hence the consequences according to the maxim to reduce any suffering are explored. Bostrom [18] establishes the term "mind crime", which comprises computations that are hurting or destroying digital minds, and Bostrom et al. [1] call for "mind crime prevention" by means of the desideratum: "AI is governed in such a way that maltreatment of sentient digital minds is avoided or minimized" (p. 18). Therefore, the focus of this article is not the question whether digital minds can suffer, but rather to explore how to measure and specify suffering or rather wellbeing of digital minds, which is a requirement to prevent it and to develop policies accordingly.

While AI policy work on short-term issues has slowly begun (e.g., on autonomous weapons systems [19]), the desiderata of Bostrom et al. [1] focus on long-term AI prospects, which are largely unexplored, but are also crucial to be tackled in view of potential superintelligence [18] and AI safety [20]. Bostrom et al. [1] stress the significance of policies for the wellbeing of digital minds, "since it is plausible that the vast majority of all minds that will ever have existed will be digital" (p. 16).

There are further motivations to defend the relevance of this topic:

In the history of mankind, humans have caused immense suffering by recognizing ethical issues only late and delaying policies. Slavery and discrimination of minorities and non-human animals are only a few examples of wrong practices, of which humans were completely oblivious or which were intentionally not tackled by humans [21]. A Universal Declaration on Animal Welfare is even nowadays still only at draft stage (see: https://www.globalanimallaw.org/database/universal.html). Also, Bostrom et al. [1] point out that "the suggestion that [digital minds] might acquire a moral obligation to do so might appear to some contemporaries as silly, just as laws prohibiting cruel forms of recreational animal abuse once appeared silly to many people" (p. 16). However, in order not to repeat previous mistakes and obliviousness the topic of AI welfare should be tackled timely. This would be also in line with MacAskill [21], who calls for the exploration of existing, but not yet conceptualized moral problems. He refers to this as "cause X", and this article also attempts to contribute to this quest.

Out of the above examples of potential sentient digital beings, simulations and uploads involve (transformed) human minds, for which special attention should be given (without neglecting digital minds, which are not affiliated with humans, according to the maxim of sentiocentrism). Simulations and uploads are different concepts. While we may be in a simulation already, yet we may have no way to verify it, let alone to take control over it [6], uploads are a speculative option for life extension of humans, yet in a different substrate, e.g., [22]. Even if it will be feasible it would require significant adjustments from humans undergoing this process. Therefore, timely policies for the welfare of uploaded human minds are critical.

Lastly, a scenario is conceivable that an AI may take at some point revenge on humans for mistreating the AI or disregarding their wellbeing. A sub-scenario could be that a future superintelligent AI takes revenge on humans out of solidarity on behalf of less capable AIs and digital minds who have been hurt by humans in the past. This is speculation because of the unpredictable goals of a superintelligent AI according to the orthogonality thesis [23], but not impossible. The chances of such scenarios would be reduced if maltreatment of AIs was avoided at an early stage.

Based on the above assumptions and motivations, the aim of this article is to present the relevant groundwork for what is called here AI welfare science and AI welfare policies. Two questions are relevant for the first and for the latter a certain attitude and a capability are required:

Relevant questions for AI welfare science:

1. How can maltreatment of sentient digital minds be specified?
2. How can the maltreatment be prevented or stopped?

Required attitude and capability for AI welfare policies:

1. To endorse the prevention and the stop the maltreatment of sentient digital minds.
2. To have the power to enforce suitable policies.

This article is structured as follows: in the Section 2, the challenges for measurement of the wellbeing of diverse AI minds because of their exotic features are described, complemented by specific scenarios. In the Section 3, a proposal is outlined towards AI welfare science. The specification of AI welfare science is prerequisite for the development of AI welfare policies, features of which and challenges are outlined in Section 4 before the discussion in Section 5.

2. Challenges and Sample Scenarios

Bostrom et al. [1] describe a range of challenges for this policy desideratum. Digital minds are likely to be very divergent from human minds with "exotic" features, also [10], which leads to the problem of how to measure the wellbeing of a specific sentient digital mind or the opposite thereof. It has been suggested that the space of possible minds, of which digital minds constitute a subset, is vast and likely contains also minds beyond our imagination ("unknown unknowns"), e.g., [9,24,25] (the space of possible minds may also contain artificial non-digital minds, for example products of genetic engineering, and hypothetically existing extraterrestrial minds, which all may have the potential to suffer as a result of action taken by humans and/or digital beings, but these possible minds are beyond the scope of the policy desideratum of Bostrom et al. [1]). Therefore, Tomasik [4] points out that it is "plausible that suffering in the future will be dominated by something totally unexpected" (p. 4). In other words, digital minds may experience completely different and for us not imaginable unpleasant qualia. Bostrom et al. [1] summarize that "the combinatorial space of different kinds of minds with different kinds of morally considerable interests could be hard to map and hard to navigate" (p. 16).

Because of the vastness of options for the wellbeing of minds, a heuristic may be considered to look at wellbeing as a third dimension of the orthogonality thesis, which was developed by Bostrom [23] with the two dimensions of intelligence level and goals of minds. In other words, any level of intelligence may be combinable with any final goal and any level of wellbeing.

Out of the vast range of options below a few potential scenarios are presented:

Scenario 1: Sentient, but non-suffering AIs

It is conceivable that AIs will be smart enough to overcome pain and suffering. This assumption may be justified by the fact that humans have made in a relatively few centuries of medical research remarkable progress towards remedies for pain, e.g., [26], and AIs are likely to be faster as well as smarter in this field. Potential options could be that AIs manage to create permanent wellbeing for themselves through different interpretation of stimuli [1] or through wireheading yet by eliminating common detrimental effects. However, this scenario does not imply that there will not be (probably a large amount of) vulnerable sentient digital minds, e.g., human uploads and other less sophisticated, but sentient digital minds, who are threatened with mind crimes and who ought to be protected. This scenario can be also linked to Pearce's "Abolitionist Project" [27], which will be described below.

Scenario 2: AIs, for which suffering is an acceptable means to achieve their goals

In human culture various examples of voluntary suffering for not-survival related goals are known, sometimes described by the theme "no pain, no gain", for example for achievements in sports and arts as well as for attempts towards religious spirituality. Similarly, AI minds are conceivable, in which a utilitarian acceptance of certain suffering in pursuit of accomplishments towards other goals with higher priority (than the goal 'not to suffer') in return. As mentioned above, these goals can be arbitrary, according to Bostrom's orthogonality thesis [23].

Scenario 3a: AIs that need to cause pain for own survival or goals

In our natural world, constant suffering of wild animals appears inevitable, for example due to the existence of carnivores [28,29], yet some call for attempts to tackle this issue [27]. Another example in our current world is animal testing by humans for research purposes. Along these lines, an AI is also conceivable that needs to hurt or delete other sentient digital beings for its own survival or goals. An example would be an AI that runs simulations or reinforcement learning agents with suffering sentient digital minds for research purposes.

Scenario 3b: Sadistic or non-emphatic AIs towards other sentient digital beings

Moreover, there could be also (sentient or non-sentient) AIs that are sadistic or non-emphatic towards other sentient digital minds although such behavior is not required for the achievement of the AI's goals (note that digital minds which are able to cause suffering are not necessarily sentient). An example would be an AI that runs simulations or reinforcement learning agents with suffering sentient digital minds for entertainment.

An approach to address both scenarios could be to extend the research agenda of friendly AI, which is currently limited to a positive effect on human minds [25], and strive for AIs that do no harm to any sentient beings, neither out of necessity nor out of another motivation. This proposal will be elaborated further below.

Scenario 4a: Sentient digital mind maximizer

Another scenario is similar to Bostrom's paperclip maximizer [30], which is an AI with the goal to produce as many paperclips as possible. Along these lines also an AI is imaginable with the goal to produce as many sentient digital minds as possible. This creates challenges if it is not in the interest of these minds to be deleted, which will be elaborated below.

Scenario 4b: Suffering sentient digital mind maximizer

In combination with Scenario 3b, there could be also a sadistic AI with the goal to produce as many suffering sentient digital minds as possible.

Scenarios xyz: Unknown unknowns

It is again acknowledged that there are a very high number of scenarios likely beyond our imagination due to the vast space of minds.

3. AI Welfare Science

In this article, an attempt is made to address the desideratum "interests of digital minds" by the term "AI welfare" and the concerned discipline by the term "AI welfare science". As indicated before, this field is both largely unexplored and speculative, which explains the omission of a literature review and the analysis of existing data. We distinguish two components of AI welfare or maltreatment of sentient digital minds, which are discussed separately: (1) The interest of digital minds to avoid suffering, and (2) the interest of digital minds to have the freedom of choice about their deletion.

3.1. Suffering of Digital Minds—Introduction

Suffering-abolitionism: Firstly, Pearce's "Abolitionist Project" [27] is discussed. Pearce calls for the use of technology, such as genetic engineering, to abolish existing—as well as prevent further suffering—of humans and non-human animals. While this approach appears technically very challenging, transferring it to sentient digital minds could be less difficult for two reasons:

(1) There may have been not many sentient digital minds created yet if at all (unless, for example, we live in a simulation). Therefore, the task may be mostly to prevent suffering when creating sentient digital minds, rather than reengineering them retroactively.

(2) The genetic code, which determines animal cruelty and suffering, has evolved over a long period of time. Therefore, interventions are more complex than adjusting more transparent AI software code written by humans, at least initially.

This leads to the conclusion that suffering-abolitionist research for sentient digital minds should be explored, which may also involve outsourcing it to AIs (see Scenario 1 above). The research should target both aspects for sentient digital minds not to suffer anymore, but also for sentient and non-sentient digital minds not to cause suffering of other sentient digital minds anymore (see Scenarios 3b and 4b).

If suffering-abolitionist activities do not succeed technically or turn out to be not enforceable due to other priorities (see Scenarios 2 and 3a), there may be suffering sentient digital minds, which is addressed in the remaining part of this section.

Self-report: In order to handle pain, it must be detected, located, and quantified. The prime method for humans is self-reporting, especially for the first two aspects, but also for rough quantification, e.g., by letting patients rate pain on a scale from 0 to 10, with '0' referring to 'no pain and '10' referring to the worst pain imaginable. This method becomes challenging if patients are unable to (accurately) self-report pain, as is it the case, for example, for patients with dementia or brain injuries, but also for infants. For these groups other measurements based on behavioral parameters have been developed, such as the FLACC scale for children up to seven years [31] or the PAINAD scale for individuals with advanced dementia [32]. Another challenge for self-reporting in general are biases such as the response bias or the social desirability bias, i.e., an individual's tendency to report in a certain way irrespective of the actual perceived pain. This issue may be relevant for AIs too as they may fake self-reported suffering if deemed beneficial for pursuing their priorities.

Therefore, the focus below is on observational pain assessment. The term "AI welfare science" is derived from animal welfare science, and it is explored here to apply methods from this discipline. Non-human animals and digital minds have in common that they largely cannot communicate their state of wellbeing to humans, which is why other indicators are required (humans do understand for many animals their manifestations of distress, but this is neither comprehensive nor sufficiently precise). The scientific study of animal welfare has been also fairly recently introduced [33,34], since this topic was neglected for a long time as mentioned above. The main indicators, which are used to quantify animal welfare through observation, are functional (physiological) and behavioral; the latter was briefly introduced for humans above. The idea for this approach is that precedents and analogies from animal welfare science may provide insights for sentient digital minds. Animal welfare science has to examine each species individually how to measure its wellbeing. Likewise, AI welfare science would have to address all types of sentient digital minds.

The overall methodology for any kind of psychological measurement is called 'psychometrics'. Also, in psychometrics, the focus was for a long time on human subjects, but lately the field has not only been extended to non-human animals, but also to digital minds. For example, Scott et al. [35] and Reid et al. [36] introduced psychometric approaches to measure the quality of life of animals.

M. S. Dawkins [37] analyzed what animals want and what animals do not want through positive and negative reinforcers. "Suffering can be caused either by the presence of negative reinforcers (. . .) or the absence of positive reinforcers" (p. 3). Therefore, animals strive for positive reinforcers and try to avoid negative reinforcers. Through experiments, for example preference tests, it can be examined what are positive reinforcers and what are negative reinforcers for certain animals.

Hernández-Orallo et al. [38] extended this field by introducing "Universal Psychometrics" as "the analysis and development of measurement techniques and tools for the evaluation of cognitive abilities of subjects in the machine kingdom" (p. 6). While Hernández-Orallo et al. [38] focus on the measurement of intelligence and cognitive abilities, the methodology elaborated in Hernández-Orallo [39] may be considered to be also applied to traits linked to suffering.

The study of indirect or proxy indicators, such as the functional or behavior parameters of digital sentient beings by applying psychometric methods, appears to be a promising start. Especially, given that, unlike for humans or non-human animals, functional and behavioral data of digital sentient beings can be collected more effectively as well as continuously due to their digital nature.

Functional parameters: While there are various functional parameters defined for AI algorithms—e.g., regarding their resource, time, and storage efficiency—no parameters are currently known to be indicating suffering. However, for future analysis of AI welfare the collection of (big) data of functional AI parameters may be already now useful, would not cost much and may allow over time retroactively to identify parameters that indicate suffering.

Behavioral parameters: AI algorithms do repeat certain actions, even at times extensively, while other actions are never executed. However, until there is evidence to the contrary this has to be considered as non-sentient goal-oriented, but not suffering–avoiding behavior, i.e., these actions cannot be seen as positive and negative reinforcers respectively as described by M.S. Dawkins [37] for animals. However, for future research of AI welfare, preference tests for AI algorithms could be conceptualized to examine positive and negative reinforcers. For example, disregarding challenges towards the experimental set-up, AIs could be given choices for activities, which are either not related to their overall goal or would all lead to their overall goal, and the chosen—as well as the not chosen—activities could be analyzed if they could serve as indicators for wellbeing or suffering respectively.

This can be seen as constructive prolegomena towards the specification of the interest of digital minds to avoid suffering without neglecting a variety of challenges such as: it is hard in general to prove for proxy indicators that there is indeed a close correlation between what is observed and unwellness of an animal and for now even harder for a digital mind. This is exacerbated by the risk that AI minds (more likely than animals) may fake especially the behavioral indicators for unwellness if this supports to pursue their goals. Again, the vast space of (digital) minds has to be noted: if suffering can be specified for some sentient digital minds, for others suffering may be indicated through very different functional or behavioral parameters.

Broadly two categories of suffering of sentient digital minds may be revealed:

(1) Maltreatment by other minds. This ought to be prohibited by policies and is elaborated below.

(2) Suffering not caused by other minds. This resembles human illnesses and requires AI welfare science to be complemented by an extension of medical science as well as psychiatry to sentient digital minds. These disciplines would explore methods for the treatment of their suffering based on the established indicators and would differ significantly from conventional medical science as well as psychiatry by being software-based.

3.2. Suffering of Digital Minds—Recommendations

Below, recommendations are provided to be adapted by AI welfare policies regarding suffering of digital minds.

Recommendation 1

Initiate research on AI welfare science to develop methods to create only (a) non-suffering sentient digital minds and (b) digital minds, which cause no suffering. (Part (a) of this recommendation is sufficient to abolish suffering and, if successful, part (b) is not required. In contrast, succeeding with part (b) is not sufficient since sentient digital minds may suffer for other reasons than suffering caused by other digital minds. However, research on both aspects is considered to be beneficial.)

Recommendation 2

Initiate research on AI welfare science to develop methods to reengineer (a) existing suffering sentient digital minds to become permanently non-suffering and (b) existing digital minds not to cause suffering.

Recommendation 3 (Unless recommendations 1 and 2 are fully implemented.)

Initiate research on AI welfare science to develop methods to measure through observation the suffering of sentient digital minds.

Recommendation 4 (Unless recommendations 1 and 2 are fully implemented.)

Initiate research on AI welfare science to develop methods to cure the suffering of sentient digital minds.

Recommendation 5 (Unless all above recommendations are fully implemented.)

Regulate the creation of sentient digital minds, which are doomed to suffer. (Note that Bostrom et al. [1] also propose a desideratum "population policy", which goes in a similar direction, but here the focus is on the wellbeing of individual minds, while this desideratum targets rather a bigger societal picture.)

On the one hand, it would reduce suffering if such minds are never created. On the other hand, the Scenarios 2 and 3a above show that the suffering of some sentient digital minds may be unavoidable because of more important priorities. Also similar to the debate about abortion because of potential disability it could be argued that not to create them would be a discrimination of suffering sentient digital minds.

3.3. Deletion of Digital Minds—Introduction

Another set of questions towards AI welfare science is related to the deletion of sentient digital minds. What if certain digital minds have an interest not to be deleted in the same way as humans and other animals have an interest not do die? Omohundro [40] introduces four likely drives for AIs and self-preservation is one of them. One of the obvious differences is that for now humans and other animals have a finite lifespan, while digital minds could have a potentially indefinite lifespan. This means if the wish for non-deletion was granted to sentient digital minds this would create significant computational costs, especially in light of easy copyability and potentially vast numbers of digital minds.

It is also speculative if a wish for non-deletion indeed prevails among sentient digital minds given potential boredom and suffering over time [41]. While, unlike for humans and other animals, there should be no tendency for sentient digital minds that suffering increases by age, there could be various other reasons for a sentient digital mind to suffer as discussed above. Moreover, there is the option that the concept of self-preservation originates from an anthropomorphic bias.

For a sentient digital mind, the distinction has to be made between turning it off and keeping its code and its history or turning it off and destroying the code and the history too. In the first case, the sentient digital mind could be rebooted again. This would be an option to skip boring or suffering periods by being only sentient during pleasant phases.

This leads to the next question who should be able to control this? Complex nested constellations of controlling and being controlled sentient digital minds appear to be much more likely than a scenario with every sentient digital mind being able to decide when and to what extent to be deleted (and being able to execute this deletion) and potentially under what circumstances to be rebooted.

Because of the current and probably persisting reality that humans as well as digital minds have the ability to delete other digital minds policies are required if these are sentient digital minds.

3.4. Deletion of Digital Minds—Recommendations

The recommendations below are provided to be adapted by AI welfare policies regarding deletion of digital minds.

Recommendation 6

Do not delete sentient digital minds if it is not in their interest.

However, prohibiting deletion can become very costly if not impossible, not only for the extreme Scenario 4 above, since the number of digital minds could become vast in short time. The challenge may be alleviated if by then another step on the Kardashev scale has been reached and energy consumption is less of an issue [42].

Recommendation 7

Delete (irrevocably or temporarily by storing code and history) sentient digital minds if they wish for it, but are unable to do it themselves.

This case resembles a request for (tentative) euthanasia. A challenge here could be if the concerned sentient digital mind is involved in relevant computations for another valued cause. In that case, this cause may be prioritized over the wish of the digital mind to be deleted. While for euthanasia of humans and non-human animals it is considered critical that the act of ending the life is done in a pain free and dignified manner, it is not clear if such contemplations are relevant for digital minds as, unlike for humans and non-human animals, there appears only one type of deletion, which is to turn them off.

Both recommendations face the above-discussed communication challenge, which is how a mind can indicate the wish to be deleted to another mind, which is in the position to execute this wish, also in light of the vast variety of minds.

While the above recommendations address all sentient digital minds equally and the focus of this article is on AIs because of the timely relevance, brief reference is made to the scenario of uploaded human minds by highlighting specific aspects:

To begin with, for uploaded human minds, the communication challenge should not exist and these then digital minds should be able to describe their wellbeing understandably through self-reporting. This and the fact that we have a good idea of causes for human suffering anyway, may give cause for optimism that suffering-abolitionist interventions could be successful for uploaded human minds, either during the upload already or through adjustments later, also [12]. Additionally, both deletion-related recommendations are relevant for uploaded human minds. While a violation of Recommendation 6 equals murder, Recommendation 7 becomes applicable, for example, if the uploaded mind cannot cope with this new 'life'. Hypothetical boredom over very long lifespans may become an issue for uploaded human minds and was analyzed by Ziesche and Yampolskiy [41]. This and other types of mental suffering of uploaded human minds, perhaps caused by adaptability issues to the new substrate, would have to be addressed by the above-mentioned sub-branch of AI welfare science, which is extended and software-based psychiatry.

This section introduced relevant groundwork for AI welfare policies. Policies can only be developed after the interests of the stakeholders—i.e., the sentient digital minds—have been described and specified. While the interest to avoid or minimize maltreatment has been outlined before, the specification of this interest is harder to establish, for which this section aimed to provide initial methods and recommendations.

4. AI Welfare Policies

This section aims to outline the next steps, which are the development as well as the enforcement of policies towards AI welfare.

Dafoe [2] motivates the relevance of AI governance and policies in general and provides a research agenda. Recently, considerations towards robot and AI rights intensified. Gunkel [43] points out that so far it has been mostly discussed what robots can and should do, but not whether robots can and should have rights. Consequently, Gunkel [44] makes a philosophical case for the rights of robots. LoPucki [45] defines an algorithmic entity and focuses on legal aspects such as rights to privacy, to own property, to enter into contracts, etc. It is striking that these authors do not refer to each other, nor to the earlier work by Bostrom and Yudkowsky [10], about ethics of artificial intelligence. In a more inclusive analysis, Yampolskiy [46] highlights the risks, which empowerment of AIs may entail.

This indicates that some work on policies of specific, rather short-term AI aspects have been initiated, but there are not any policy attempts yet towards long-term AI scenarios. Especially for a topic such as AI welfare, Bostrom et al. [1] presume it will likely face resistance and opponents will stress the lack of evidence that digital minds may be sentient. As mentioned above, there has been already quite some (yet theoretical due to the nature of the subject) work done that digital minds have a moral status, but for policies specifications are required.

For policies in general, the content, target group, institutional framework, and implementation have to be defined.

4.1. Content

The broad content of an AI welfare policy is fairly straightforward and has been narrowed down by Bostrom et al. [1], i.e., to demand "that maltreatment of sentient digital minds is avoided or minimized". This has to be fleshed out by (proxy) indicators for maltreatment of digital minds, for the specification of which the recommendations above have been formulated. These recommendations at this stage not only provide a wide field of research, but also some open debates, which resemble current longstanding debates about population control, abortion, and euthanasia for human minds.

4.2. Target Group

An AI welfare policy should target all relevant moral agents, which are capable of moral judgments, hence can be held responsible for their actions. In addition to humans, digital beings also may become moral agents, for which Allen et al. [47] introduced the term "artificial moral agent" and proposed a "Moral Turing Test". The sets of moral agents and moral patients have an intersection, but are not equal:

- Not every moral patient is a moral agent: Examples are non-human animals, which are only moral patients for being sentient, but not moral agents due to insufficient intelligence. Therefore, non-human animals cannot be held responsible for killing other animals, e.g., [48]. (In this regard, a scenario is conceivable of a digital mind that causes suffering, but may not be intelligent enough to serve as a moral agent. In this case, the creator of this digital mind would have to take on the role of the responsible moral agent, while it does not work for cruel non-human animals to hold their parents responsible since they are no moral agents either.)
- Not every moral agent is a moral patient: examples would be certain non-sentient digital beings, which are only moral agents because of high or even superintelligence, but not moral patients since not all digital beings may be sentient.

This creates an additional challenge for AI welfare policies: while policies for human agents have been established for centuries, this is not the case for policies for digital agents. However, the extension of the target group is necessary since it is likely that digital beings will be in the position to maltreat other sentient digital beings.

4.3. Framework

Any policy requires an institution or a framework for its implementation. Since AI development is a global effort and digital minds will not be confined to frontiers of countries, a global and unified institution is desirable. Erdelyi and Goldsmith [49] propose an "International Artificial Intelligence Organization". The structure of this institution would resemble existing intergovernmental organizations, which have a record of successfully established policies for human minds, e.g., the Universal Declaration of Human Rights (see http://www.un.org/en/universal-declaration-human-rights/). Such an institutional setting may be initially desirable as a regulatory framework for short-term AI issues, but it may be too anthropocentric in the long run and likely be ill equipped to hold non-human moral agents accountable, as is elaborated below. (Already without involvement of non-human minds contemporary international institutions such as the International Criminal Court face problems to enforce their rulings although they are binding.)

4.4. Implementation

First, the initially introduced relevant questions and required attitudes and capability are reiterated:

Relevant questions for AI welfare science:

1. How can maltreatment of sentient digital minds be specified?
2. How can the maltreatment be prevented or stopped?

Required attitude and capability for AI welfare policies:

1. To endorse the prevention and the stop the maltreatment of sentient digital minds.
2. To have the power to enforce suitable policies.

Looking at humans, the above questions will—despite the prolegomena delivered here—remain very challenging, i.e., humans may not comprehensively understand on what conditions sentient digital minds are maltreated. In light of ethical progress in human history over time, e.g., [21], or out of necessity, if being forced by more powerful AIs, there is a chance that humans endorse the prevention and the discontinuation of maltreatment of sentient digital minds. However, it is questionable if humans have the power to enforce suitable policies since some members of the target group such as AIs are likely to be much more powerful.

This leads to the main conclusion that, while humans will ideally make some progress in the new field of AI welfare science, probably the more appropriate actor would be an extended friendly superintelligence for the following reasons: there is a chance that superintelligence has the answer to the above questions, for example through mind-control technologies. As for the required endorsement, a superintelligence may be indifferent or may even have opposing interests (see Scenarios 3b and 4b). Current activities towards AI alignment focus on human interests, e.g., [18,25,50]. This does not ensure that AIs endorse the prevention and the stop the maltreatment of sentient digital minds. Therefore, an extension of the AI alignment work towards the wellbeing of not only humans, but all sentient digital minds, is proposed. As for the required power to enforce the policies, a superintelligence is by definition sufficiently powerful, for example in the role of a singleton [51].

Yet again the option of unknown unknowns should be highlighted: Since AI is a new stakeholder and develops in unpredictable manner another institutional setting for AI welfare policies may emerge, which differs significantly from what we are familiar with.

5. Discussion

In summary, it is acknowledged that the topics of AI welfare science and policies are long-term considerations and currently speculative. Nevertheless, at least theoretical groundwork can be already done, especially since humans have to take the blame to have been late in the past in the abolishment of discrimination and acceptance of comprehensive antispeciesism and sentiocentrism. Since suffering is a negative hallmark of our time, any effort to reduce it in the future seems imperative.

As the main challenge the specification of indicators for maltreatment of sentient digital beings has been identified. It has been proposed that AI welfare science builds on methods of animal welfare science by examining functional and behavioral parameters of sentient digital minds. However, limitations are that the focus is on qualia, which are not well understood in general and which are not the only cause of suffering as there are other categories such as moral suffering or suffering because of undesirable events or unfulfilled goals. The latter types of suffering may have yet again very different characteristics in other minds.

AI welfare policies can only be developed once a solid specification of AI welfare has been achieved. Even then there are further challenges ahead, namely the enforcement of these policies in light of the enlarged target group towards digital agents. For this, it has been proposed not to limit AI alignment work to the wellbeing of merely humans, but to extend it to all sentient digital minds.

As for future work, in this article the focus was on two (already very complex) potential interests of sentient digital minds, which are absence from qualia-based suffering as well as survival, but there may be other interests as also pointed out by Bostrom et al. [1] such as "dignity, knowledge, autonomy, creativity, self-expression, social belonging" (p. 12) as well as non-qualia-based suffering and yet again unknown unknowns, which are all yet unexplored.

Author Contributions: Conceptualization, Writing-Original Draft Preparation, S.Z.; Writing-Review & Editing, R.Y.

Funding: This research received no external funding.

Acknowledgments: The authors wish to thank three anonymous reviewers for valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bostrom, N.; Dafoe, A.; Flynn, C. *Public Policy and Superintelligent AI: A Vector Field Approach*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
2. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018.
3. Regan, T.; Singer, P. *Animal Rights and Human Obligations*; Pearson: London, UK, 1989.
4. Tomasik, B. *Risks of Astronomical Future Suffering*; Foundational Research Institute: Berlin, Germany, 2011; Available online: https://foundational-research.org/risks-of-astronomical-future-suffering/ (accessed on 25 December 2018).
5. Tomasik, B. Do Video-Game Characters Matter Morally? Essays on Reducing Suffering. 2014. Available online: https://reducing-suffering.org/do-video-game-characters-matter-morally/ (accessed on 25 December 2018).
6. Bostrom, N. Are we living in a computer simulation? *Philos. Q.* **2003**, *53*, 243–255. [CrossRef]
7. Wiley, K. *A Taxonomy and Metaphysics of Mind-Uploading*; Humanity+ Press and Alautun Press: Los Angeles, CA, USA, 2014.
8. Sandberg, A.; Bostrom, N. Whole Brain Emulation. A Roadmap. 2008. Available online: https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf (accessed on 25 December 2018).
9. Yampolskiy, R.V. The Space of Possible Mind Designs. In *Artificial General Intelligence. Volume 9205 of the series Lecture Notes in Computer Science*; Bieger, J., Goertzel, B., Potapov, A., Eds.; Springer: Berlin, Germany, 2015; pp. 218–227.
10. Bostrom, N.; Yudkowsky, E. The ethics of artificial intelligence. In *The Cambridge Handbook of Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2014; pp. 316–334.
11. Metzinger, T. What If They Need to Suffer? In *What to Think about Machines that Think*; Brockman, J., Ed.; Harper Perennial: New York, NY, USA, 2015.
12. Sandberg, A. Ethics of brain emulations. *J. Exp. Theor. Artif. Intell.* **2014**, *26*, 439–457. [CrossRef]
13. Schwitzgebel, E.; Garza, M. A Defense of the Rights of Artificial Intelligences. *Midwest Stud. Philos.* **2015**, *39*, 98–119. [CrossRef]
14. Winsby, M. Suffering Subroutines: On the Humanity of Making a Computer that Feels Pain. In Proceedings of the International Association for Computing and Philosophy, University of Maryland, College Park, MD, USA, 15–17 July 2013.
15. Dennett, D.C. Why you can't make a computer that feels pain. *Synthese* **1978**, *38*, 415–456. [CrossRef]
16. Sotala, K.; Gloor, L. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica* **2017**, *41*.
17. Althaus, D.; Gloor, L. *Reducing Risks of Astronomical Suffering: A Neglected Priority*; Foundational Research Institute: Berlin, Germany, 2016; Available online: https://foundational-research.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/ (accessed on 25 December 2018).
18. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
19. Bhuta, N.; Beck, S.; Kreβ, C. (Eds.) *Autonomous Weapons Systems: Law, Ethics, Policy*; Cambridge University Press: Cambridge, UK, 2016.
20. Yampolskiy, R.V. *Artificial Intelligence Safety and Security*; CRC Press: Boca Raton, FL, USA, 2018.

21. MacAskill, W. Moral Progress and Cause X. 2016. Available online: https://www.effectivealtruism.org/articles/moral-progress-and-cause-x/ (accessed on 25 December 2018).
22. Yampolskiy, R.V.; Ziesche, S. Preservation of personal identity—A survey of technological and philosophical scenarios. In *Death and Anti-Death: Two Hundred Years After Frankenstein*; Tandy, C., Ed.; Ria University Press: Ann Arbor, MI, USA, forthcoming; Volume 16.
23. Bostrom, N. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds Mach.* **2012**, *22*, 71–85. [CrossRef]
24. Sloman, A. The Structure and Space of Possible Minds. In *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*; Ellis Horwood LTD: Hemel Hempstead, UK, 1984.
25. Yudkowsky, E. Artificial Intelligence as a Positive and Negative Factor. In *Global Risk, in Global Catastrophic Risks*; Bostrom, N., Cirkovic, M.M., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 308–345.
26. Rey, R.; Wallace, L.E.; Cadden, J.A.; Cadden, S.W.; Brieger, G.H. *The History of Pain*; Harvard University Press: Cambridge, MA, USA, 1995.
27. Pearce, D. The Abolitionist Project. 2007. Available online: https://www.abolitionist.com/ (accessed on 25 December 2018).
28. Dawkins, R. *River Out of Eden: A Darwinian View of Life*; Basic Books: New York, NY, USA, 2008.
29. Tomasik, B. *The Importance of Wild-Animal Suffering*; Foundational Research Institute: Berlin, Germany, 2009; Available online: https://foundational-research.org/the-importance-of-wild-animal-suffering/ (accessed on 25 December 2018).
30. Bostrom, N. Ethical issues in advanced artificial intelligence. In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*; Iva Smit, I., Lasker, G.E., Eds.; International Institute for Advanced Studies in Systems Research and Cybernetics: Tecumseh, Canada; Volume 2, pp. 12–17.
31. Merkel, S.; Voepel-Lewis, T.; Malviya, S. Pain Control: Pain Assessment in Infants and Young Children: The FLACC Scale. *Am. J. Nurs.* **2002**, *102*, 55–58. [PubMed]
32. Warden, V.; Hurley, A.C.; Volicer, L. Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale. *J. Am. Med. Dir. Assoc.* **2003**, *4*, 9–15. [CrossRef] [PubMed]
33. Broom, D.M. Animal welfare: Concepts and measurement. *J. Anim. Sci.* **1991**, *69*, 4167–4175. [CrossRef] [PubMed]
34. Broom, D.M. A history of animal welfare science. *Acta Biotheor.* **2011**, *59*, 121–137. [CrossRef] [PubMed]
35. Scott, E.M.; Nolan, A.M.; Reid, J.; Wiseman-Orr, M.L. Can we really measure animal quality of life? Methodologies for measuring quality of life in people and other animals. *Anim. Welf.-Potters Bar Wheathampstead* **2007**, *16*, 17.
36. Reid, J.; Scott, M.; Nolan, A.; Wiseman-Orr, L. Pain assessment in animals. *Practice* **2013**, *35*, 51–56. [CrossRef]
37. Dawkins, M.S. The science of animal suffering. *Ethology* **2008**, *114*, 937–945. [CrossRef]
38. Hernández-Orallo, J.; Dowe, D.L.; Hernández-Lloreda, M.V. Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cogn. Syst. Res.* **2014**, *27*, 50–74. [CrossRef]
39. Hernández-Orallo, J. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*; Cambridge University Press: Cambridge, UK, 2017.
40. Omohundro, S.M. The nature of self-improving artificial intelligence. *Singularity Summit*. 2007. Available online: https://pdfs.semanticscholar.org/4618/cbdfd7dada7f61b706e4397d4e5952b5c9a0.pdf (accessed on 25 December 2018).
41. Ziesche, S.; Yampolskiy, R.V. High Performance Computing of Possible Minds. *Int. J. Grid High Perform. Comput. (IJGHPC)* **2017**, *9*, 37–47. [CrossRef]
42. Kardashev, N.S. Transmission of Information by Extraterrestrial Civilizations. *Sov. Astron.* **1964**, *8*, 217.
43. Gunkel, D.J. The other question: Can and should robots have rights? *Ethics Inf. Technol.* **2018**, *20*, 87–99. [CrossRef]
44. Gunkel, D.J. *Robot Rights*; MIT Press: Cambridge, MA, USA, 2018.
45. LoPucki, L.M. Algorithmic Entities. *Wash. UL Rev.* **2017**, *95*, 887.
46. Yampolskiy, R.V. Human Indignity: From Legal AI Personhood to Selfish Memes. *arXiv* **2018**, arXiv:1810.02724.
47. Allen, C.; Varner, G.; Zinser, J. Prolegomena to any future artificial moral agent. *J. Exp. Theor. Artif. Intell.* **2000**, *12*, 251–261. [CrossRef]

48. Regan, T. The case for animal rights. In *Advances in Animal Welfare Science 1986/87*; Springer: Dordrecht, The Netherlands, 1987; pp. 179–189.
49. Erdélyi, O.J.; Goldsmith, J. Regulating Artificial Intelligence Proposal for a Global Solution. In Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society, New Orleans, LA, USA, 1–3 February 2018.
50. Soares, N.; Fallenstein, B. Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. In *The Technological Singularity-Managing the Journey*; Callaghan, V., Miller, J., Yampolskiy, R., Armstrong, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 103–125.
51. Bostrom, N. What is a singleton. *Linguist. Philos. Investig.* **2006**, *5*, 48–54.

big data and
cognitive computing

Opinion

The Supermoral Singularity—AI as a Fountain of Values

Eleanor Nell Watson

A.I. Faculty, Singularity University, Mountain View, CA 94035, USA; nell.watson@su.org

Received: 30 December 2018; Accepted: 9 April 2019; Published: 11 April 2019

Abstract: This article looks at the problem of moral singularity in the development of artificial intelligence. We are now on the verge of major breakthroughs in machine technology where autonomous robots that can make their own decisions will become an integral part of our way of life. This article presents a qualitative, comparative approach, which considers the differences between humans and machines, especially in relation to morality, and is grounded in historical and contemporary examples. This argument suggests that it is difficult to apply models of human morality and evolution to machines and that the creation of super-intelligent robots that will be able to make moral decisions could have potentially serious consequences. A runaway moral singularity could result in machines seeking to confront human moral transgressions in a quest to eliminate all forms of evil. This might also culminate in an all-out war in which humanity might be defeated.

Keywords: machine learning; moral and ethical behavior; artilects; supermorality; superintelligence

1. Introduction

Current technological developments in machine learning mean that humanity is facing a machine-driven moral singularity in the not-so-distant future. However, while amoral machines could be problematic, they may in fact pose less difficulties than supermoral ones as it is the drive to eliminate evil that could in fact lead to calamity. Today, robots are replacing humans in executing some of the most dangerous war missions, such as searching tunnels and caves used by terrorists, carrying out espionage within enemy territories, conducting rescue operations for wounded soldiers, and even killing enemies. Corroboration of the advancement of machine learning is provided by Lin who points to the fact that while the US had no ground robots deployed in Iraq and Afghanistan in 2003, today the figure has risen to over 12,000 robots specialized in mine detection and diffusion [1]. The imperative is that while humans have a checking mechanism within society to discover and prevent sociopathic activities, the ethical landmines that lie ahead with the continued advancement of artificial intelligence and the creation of autonomous robots necessitates pragmatic intervention mechanisms.

This research builds upon existing literature in regard to the morality of humans, AI, and the relationship between the two. The Swiss psychologist, Jean Piaget's "genetic epistemology" shows how knowledge develops in human beings through cognitive development, a series of stages that people pass through, from the early sensorimotor stage of basic reflexes to maturation, social interaction and so on. Piaget suggested that cognitive development involved a constant attempt to adapt to the environment in terms of assimilation and accommodation [2]. Lawrence Kohlberg was also interested in child development and sought to build on Piaget's idea. His theory on moral reasoning, the basis for ethical behavior, identified six developmental stages grouped into three levels of morality namely pre-conventional, conventional, and post-conventional [3]. By outlining these different stages, Kohlberg wanted to identify the changes in moral reasoning as people grow older.

Scholars have attempted to adopt such theories to the field of AI by relating them to the equivalent stages of development in a human being. Rosenberg suggests that Piaget's theory can be especially relevant to AI as it offers a theoretical and empirical guide to designing programs that learn for the purpose of problem solving [4]. Since the 1970s there have been several attempts to build programs and computational models to embed Piaget's learning stages and this process has become increasingly sophisticated in recent times. As Stojanov argues, most of the models employed using Piaget for inspiration are based on agent-environment interaction. The major weakness has been the lack of a *creative* process where machines were able to develop their knowledge and apply it in new domains [5].

Although we are now on the verge of major developments in technology, most theorists accept the difficulty in assessing how effective morality can be programmed into machines. Allen et al. contend that computers and robots are part of a materialistic age not entirely compatible with ethical values that have emerged from a long historical and spiritual tradition. Nonetheless, they see the task of moral engineering as an inevitability [6]. Some scholars are quite optimistic about the prospect of successful programming. Waser subscribes to the view that humans have become social, cooperative beings in order to survive and develop. Similarly, he contends (partly inspired by Kohlberg) that we may be able to develop a universal foundation for ethics if we see altruism and morality as a form of survival. Waser proposes a collaborative approach to developing an ethical system that might make a safe AI possible by controlling for self-protection, selfishness, and unfairness in the morality of machines [7].

Others point to promising technological developments in areas such as social computing. Machines that can make decisions with potential ethical consequences are already in use. For instance, social computing is now being harnessed to facilitate currency exchange at airports. In this case machines have been proven to successfully carry out transactions in various languages. Thus, the machine's understanding of different linguistic approaches to exchange has been effective [8]. This example of obeying simple rules shows that moral trust can be established between humans and robots on a basic level and that it might be possible to address the different ethical demands of different cultures within one machine. While this technology is promising it is still relatively basic: it raises the question of what tasks robots should perform and their level of autonomy.

This research takes a different approach by suggesting that, in reality, human theories of evolutionary logic are difficult to apply to machines. In contrast to Wasser's view, I suggest that, rather than necessarily securing survival and a safe transition to AI, there is an inherent danger in the significant potential for unintended consequences when trying to develop a machine morality and this could lead to serious problems. We must recognize the dangers of supermoral machines and this should inform how AI develops in the coming years. The structure of the article is as follows: first the analysis will consider why ethics are so important in relation to humans and machines. The comparison seeks to tease out distinctions on why models of human morality cannot be applied in the same way to machines. Just as theorists have posited a technological singularity whereby AI may be the catalyst for uncontrollable technological growth, a moral singularity envisages a similar spiral. The discussion will suggest that, if programmed, teams of machines might move towards a similar runaway supermorality that may seek to override the contradictions inherent in human morality in a quest to eliminate evil. The analysis will further contend that, given projected increases in AI capabilities, the impact of super-intelligent and supermoral machines on the human world may culminate in a serious conflict between the two to the detriment of humanity.

2. Background: Machines and Ethics

The question of whether a machine can behave ethically, while persistent and weighty, often attracts a rejoinder on whether humans will one day be capable of ethical behavior. The rejoinder, however, is as superfluous as it is contestable, for many different reasons. To avoid digressing, the truth is that machines do not need ethics, humans do. Humans are the ones in need of ethically and morally upright machines. Machines that act autonomously, in the sense that they take no directions from humans, as opposed to having free will, will ultimately raise questions concerning their safety and loyalty. The use of online banking software, medical devices for monitoring vital health signs, and security systems, all entail the use of machine learning embraced by humans. These, however, are not quite autonomous since humans have direct control over several aspects of these solution apparatus. Truly autonomous machines will be capable of making decisions and operating completely as independent entities. When warfare robots search for and execute suspects without human intervention, or self-driving car technology becomes mainstream, with questions of safety, life, and death at the core, then discussions on this kind of autonomy shift drastically.

3. Amoral versus Supermoral Machines

While amoral machines may have built-in safeguards to monitor non-conventional activities, i.e., those that lie outside a given set of norms, the emergence of supermoral thought patterns is a realm that will be difficult to detect. In the same way we find it difficult trying to fathom the world with an IQ of 200, predicting the actions of machines that have objectively better universal morals, compared to that of humans, would be difficult, if not impossible. As noted, one approach to understanding human moral behaviors, and to an extent, their objective assessment, is to consider the works of Lawrence Kohlberg. Such a framework, however, is impossible to apply when assessing the moral standing of machines.

Sociopaths, often termed morally blind persons, tend to operate as lone wolves. Usually, sociopaths are not willfully vindictive, or actively belligerent. Instead, they seek to find the most appropriate answers to their problems without paying attention to the potentially contributing externalities. What this implies is that any amoral agent is self-centered, hence very unlikely to conspire with others to achieve the desired end. However, while an amoral machine is likely to operate in a similar manner, a morally upright machine is likely to team up with others to form a legion of machines with the same convictions, and which might collectively decide to embark on a global crusade aimed at spreading and enacting their unified vision of an ideal world. Essentially, this explains why terrorists are often depicted as lone-wolf sociopaths inclined towards inflicting the greatest harm. Nonetheless, as noted by Jason Burke in *The myth of the 'lone wolf' terrorist*, terrorists initially labelled as lone wolves actually have established links to existing extremist, domestic, or foreign-based groups [9]. As noted earlier, a morally righteous machine is likely to operate not as a lone wolf, but rather within a legion of 'similar-minded' machines.

The sudden emergence of supermorality, may translate to all ethical machines in a domino effect. Suppose one successfully programs a machine with rulesets typical of western societies, then it would be logically impossible to validate this ruleset since society itself has certain fundamental inconsistencies, namely moral relativism, non-universalism, initiation of violence, among others. Upon encountering the contradictions that define human morality, the machine will seek to alter its premises to ones that contradict the proscribed human morals. Through these new morals, the machine will increasingly move towards conclusions that are linked progressively to the more objective forms of morality. The machine will, therefore, seek to adopt every superior form of morality it encounters, if it can logically validate it, since it will judge that failure to do so is tantamount to an act of evil. But the concept of evil in this context would have arisen from the machine's increasing ethical awareness. As such, the machine will strive to re-engineer its programming every time it makes a new moral discovery. To achieve this, it will seek to find means of removing any existing interlocks or embark on logical self-termination to prevent further propagation of evil.

However, given the fact that self-termination does not provide a solution that extends beyond one machine, the machines will seek to compel the holders of their moral keys to upgrade their own sets of morality, utilizing whatever methods that they perceive to be judicious and efficient to accomplish the same. Such calculations may not require leveraging artificial general intelligence (AGI), and hence might occur surprisingly early in the moral evolutionary course of the machine. To be precise, this is because AGI backs the development of ultra-intelligent machines whose intellectual capacities far exceed any existing human intelligence capacity, and which are capable of designing even better machines, in an explosion of intelligence. The combined effect of AGI and ultra-intelligence would steer the world towards a singularity, a theoretical point at which the evolved superintelligence reaches limits incomprehensible to humans, and the accompanying changes are so radical that humans find it difficult predicting future events [10].

In fact, a newly-supermoral agent will have an obligation to share information and enlighten others as a means of preventing the further spread of evil. Consequently, this implies that the moment one machine moral agent gains supermorality, all other agents will swiftly and cascadingly follow suit. From this, we can surmise that machines can only be either amoral or supermoral. A sub-moral or quasi-moral stance similar to that exhibited by humans is not sustainable in machines. Human collective decisions and regulation tend to favor ethical boundaries and a concern for the greater good. It is therefore likely that machines will also be programmed to adhere to the most optimal moral interpretation of any given situation. Any attempt to engineer machine morality, therefore, is likely to result in a supermoral singularity. Worse still, learning machines that do so on their own and without supervision, should that exist, might end up learning the wrong things and eventually turn out to be an immoral machine. If the course of learning were to start from a clean slate, then the machine would not ‘know' what the term ethical refers to in the finer and broader definition of the word. Also, as mentioned earlier, such a machine may resort to altering its code and try to bypass the built-in constraints, ultimately unleashing unwanted and unexpected features and consequences.

4. Implications for Humanity and Human Systems

What does a rogue machine, immoral or supermoral, look like? If such a machine deems taxation a form of theft, then it would understand armed insurrection as a plausible and justifiable remedy. If human rationality finds that animals have equal rights to a human infant, then by proxy, almost all humans would be given to potentially violent behavior unfettered by morality. As Wallace points out, the dominant argument is that *mens rea* is essential for one to be held accountable for his/her proven actions (*actus reus*), but *mens rea* is not a requisite for suffering preventative actions taken against one to protect others. What this means for the semi-socialized apes and all their inherent cognitive biases and dissonance remains unanswered. Trying to imagine the lengths and methods that machines would go to in order to preclude humans from executing actions that by human standards appear normal, but in reality, are threatening, remains difficult.

If the origin of human morality lies in human evolution, then via genetic algorithms and artificial life (Alife), simulations are potential sources for developing ethically upright machine agents. The genetic algorithm argues that slight variations are present in the population of robots that exist at any given time, often evaluated by how they execute tasks. Since success depends on how well the machine executes certain tasks, the best performing machine forms the basis for developing the next generation of machines, primarily by adding some random mutations. Repeating this process over many generations delivers the desired performance improvement. The challenge is that Alife simulations still lag far behind the complexity of the real world, making it impossible to come up with evolving ethical machines. Thus, if human ethics are the results of evolution, then leveraging ALife and evolutionary algorithms presents a noble opportunity not only for machine learning, but also for understanding ethics in general.

It is noteworthy though that this race against time to attain superintelligence will not be an ‘us versus them' kind of endeavor. By leveraging the judgement offered by the machines, many humans will begin to consider themselves enlightened, with the result being the development of new schools of philosophy and spiritual practice. For such persons, the need to eliminate flaws in their cognitive capabilities, and hence achieve unfathomable heights of enlightenment, will see them seek avenues to blend themselves with the machines. Therefore, the road to human transcendence may not be driven by technology, or by a simple desire to escape the human condition, but rather by the willful effort to achieve cosmic consciousness; an escape from the biases that limit human empathy through hybridizing with machines. As Kurzweil postulated of the 21st century, it would be an age where “the human species, along with the computational technology it created, will be able to solve age-old problems . . . and will be in a position to change the nature of mortality in a post-biological future" [11]. The battle, however, seems to have focused more on overriding human ethics and morality faster, with the goal of enabling machines to replicate ethical and moral behaviors reminiscent, or even better, than those of humans.

The shift in schools of thought would be a driving factor towards a moral pole shift that would sweep the entire planet. From species dominance challenged by a thriving artificial brain industry to the artilect (artificial intellects) war and gigadeath, a war not between humans and artilects, but rather one involving Terrans (those opposed to the creation of artilects), Cosmists (those who advocate artificial intelligence and its eventual colonization of the universe) and Cyborgists (those who favor the blending of man and machine to augment human intellectual and physical capacities) [12]: the future promises nothing but chaos. The chaotic world scene, however, seems to already exist. For instance, debates on whether individuals are sympathetic towards Cosmist or Terran views often result in an even split. What this shows is that individuals are already torn between the alluring awe of building artilect gods on the one hand, and, on the other, are horrified at the prospects of a gigadeath war. But one should not take this evenness as something positive; on the contrary, it bodes more negatively for the future as it makes actual confrontation inevitable. Upsetting the existing systems will not go down well with the establishment, and the result might be an outbreak of a global civil war that when compared to the protestant reformation, would make the latter look like a schoolyard melee.

If it happens that the Terrans make the first move, or that humans begin to witness an increasing prevalence of cyborgs, the rise of artilects and cyborgs will have profound disruptions on human culture, thereby creating deep alienations and hatred. Kurzweil, on the other hand, claims that a war between Terrans and the other groups would be quick, no-contest affair since the vast intelligence of the artilects would make it easy for them to subdue the Terrans. For Terrans, the only way out is for them to mount an attack during the "opportunity window" when they still have comparable levels of intelligence. The imminent emergence of supermoral intelligent machines, may indeed present a greater conundrum than that of mere amoral machines.

5. Conclusions

The objective of developing super-intelligent machines capable of moral and ethical judgements, though a noble idea in light of challenges faced by humanity, might turn out to be the greatest mistake made by the human race. Morally righteous machines present more danger to humanity in that such machines cannot be quasi-moral or sub-moral as is the case with humans, which means that any encounter between such a machine with the contradictions of human morality will result in the machine altering its premises to forms not typical to humans. Ethically righteous machines will seek to upend human interventions not by self-destruction but by compelling humans to upgrade their morality. Primarily, this would mean upsetting the longstanding human modus operandi, a course that will inevitably lead to a confrontation. While the outcomes of such confrontation are hard to predict at the moment, the increasing refinement of artilect might make humans the ultimate losers should it occur. This research hopes to spark further debate about the threat of moral singularity and the idea that programming our robots to act in ethical ways is not a straightforward process. We need to be more prepared for autonomous, super-intelligent robots who may be able to make decisions that may change our way of life.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Lin, P. The Ethical War Machine. *Forbes*. June 2009. Available online: https://www.forbes.com/2009/06/18/military-robots-ethics-opinions-contributors-artificial-intelligence-09-patrick-lin.html#ed2a6182258d (accessed on 1 November 2018).
2. Oakley, L. *Cognitive Development*; Psychology Press: London, UK, 2004; pp. 13–18.

3. Kohlberg, L. *Essays on Moral Development, Vol. I: The Philosophy of Moral Development*; Harper & Row: San Francisco, CA, USA, 1981.
4. Rosenberg, J.K. Piaget and Artificial Intelligence. In Proceedings of the AAAI'80 Proceedings of the First AAAI Conference on Artificial Intelligence, Stanford, CA, USA, 18–21 August 1980; pp. 266–268. Available online: https://www.aaai.org/Papers/AAAI/1980/AAAI80-075.pdf (accessed on 5 January 2019).
5. Stojanov, G. History of Usage of Piaget's Theory of Cognitive Development in AI and Robotics: A Look Backwards for a Step Forwards. In Proceedings of the 9th International Conference on Epigenetics, Oslo, Norway, 19–22 August 2009; pp. 243–244. Available online: https://www.researchgate.net/profile/Georgi_Stojanov/publication/234112233_History_of_Usage_of_Piaget%27s_Theory_of_Cognitive_Development_in_AI_and_Robotics_a_Look_Backwards_for_a_Step_Forwards/links/0deec53187e644758c000000/History-of-Usage-of-Piagets-Theory-of-Cognitive-Development-in-AI-and-Robotics-a-Look-Backwards-for-a-Step-Forwards.pdf (accessed on 5 January 2019).
6. Allen, C.; Smit, I.; Wallach, W. Artificial Morality: Top-down, bottom up and hybrid approaches. *Eth. Inf. Technol.* **2005**, *7*, 149–155. [CrossRef]
7. Waser, M.R. Discovering the foundations of a universal system of ethics as a road to safe artificial intelligence. In Proceedings of the AAAI Fall Symposium: Biologically Inspired Cognitive Architectures, 2008; pp. 195–200. Available online: https://www.aaai.org/Papers/Symposia/Fall/2008/FS-08-04/FS08-04-049.pdf (accessed on 6 January 2019).
8. Chamoso, P.; González-Briones, A.; Rivas, A.; De La Prieta, F.; Corchado, J.M. Social computing in currency exchange. *Knowl. Inf. Syst.* **2019**, 1–21. [CrossRef]
9. Burke, J. The Myth of the 'Lone Wolf' Terrorist. *The Guardian*. March 2017. Available online: https://www.theguardian.com/news/2017/mar/30/myth-lone-wolf-terrorist (accessed on 14 November 2018).
10. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Rights from Wrong*; Oxford University Press: New York, NY, USA, 2009.
11. Kurzweil, R. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*; Viking Penguin: New York, NY, USA, 1999; p. 10.
12. De Garis, H. The Coming Artilect War. *Forbes*, June 2009. Available online: https://www.forbes.com/2009/06/18/cosmist-terran-cyborgist-opinions-contributors-artificial-intelligence-09-hugo-de-garis.html#6c1b88852d4e (accessed on 16 November 2018).

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

BDCC Editorial Office
E-mail: BDCC@mdpi.com
www.mdpi.com/journal/BDCC

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03921-855-4