How Feasible Is the Rapid Development of Artificial Superintelligence?

Kaj Sotala, Foundational Research Institute

Abstract. What kinds of fundamental limits are there in how capable artificial intelligence (AI) systems might become? Two questions in particular are of interest: 1) How much more capable could AI become relative to humans, and 2) how easily could superhuman capability be acquired? To answer these questions, we will consider the literature on human expertise and intelligence, discuss its relevance for AI, and consider how AI could improve on humans in two major aspects of thought and expertise, namely simulation and pattern recognition. We find that although there are very real limits to prediction, it seems like AI could still substantially improve on human intelligence.

Introduction

Since Turing (1950), the dream of artificial intelligence (AI) research has been the creation of a "machine that could think". While current expert consensus believes the creation of such a system to still take several decades if not more (Müller & Bostrom 2016), recent progress in AI has still raised worries about the challenges involved with increasingly capable AI systems (Future of Life Institute 2015, Amodei et al. 2016).

In addition to the risks posed by near-term developments, there is the possibility of AI systems eventually reaching superhuman levels of intelligence, eventually breaking out of human control (Bostrom 2014). Various research agendas and lists of research priorities have been suggested for managing the challenges that this level of capability would pose to society (Soares & Fallenstein 2014, Russell et al. 2015, Amodei et al. 2016, Taylor et al. 2016).

For managing the challenges presented by increasingly capable AI systems, one needs to know how capable those systems might ultimately become, and how quickly. If AI systems can rapidly achieve strong capabilities, becoming powerful enough to take control of the world before any human can react, then that implies a very different approach than one where AI capabilities develop gradually over many decades, never getting substantially past the human level (Sotala & Yampolskiy, 2015). We might phrase these questions as:

1. How much more capable can AIs become relative to humans?
2. How easily (in terms of time and resources required) could superhuman capability be acquired?

Views on these questions vary. Authors such as Bostrom (2014) and Yudkowsky (2008) argue for the possibility of a fast leap in intelligence, with both offering hypothetical example scenarios where AI rapidly acquires a dominant position over humanity. On the other hand, Anderson (2010) and Lawrence (2016) appeal to fundamental limits on predictability – and thus intelligence – posed by the complexity of the environment.

The argument for limits of intelligence (Anderson 2010, Lawrence 2016) could be summarized as saying that, past a certain point, increased intelligence is only of limited benefit, for the unpredictability of the environment means that you would have to spend exponentially more resources to evaluate a vastly increasing amount of possibilities.

Noise also accumulates over time, reducing the reliability of your models. For many kinds of predictions, increasing the prediction window would require an exponential increase in the amount of measurements (Martela 2016). For instance, weather models become increasingly uncertain when projected farther out in time. Forecasters can only access a limited amount of observations relative to the weather system's degrees of freedom, and any initial imprecisions will magnify over time and cause the accuracy to deteriorate (Buizza, 2002). In general, the accuracy of any long-term prediction will be limited by data uncertainty, model uncertainty, and the available computational time. Similar considerations would also apply to attempts to predict things such as the behavior of human societies. The advantage that even a superhuman intelligence might have over humans may be limited.

On the other hand, it is not obvious whether this point of view really is in conflict with the assumption of AI being able to quickly grow to become powerful. There being limits to prediction does not imply that humans would be particularly close to the limits, nor that it would necessarily take a great amount of time to move from sub-human to superhuman capability.

This article attempts to consider these questions by considering what we know about expertise and intelligence. After reviewing the relevant research on human expertise, we will discuss its relevance for AI, and consider how AI could improve on humans in two major aspects of thought and expertise, namely simulation and pattern recognition. Our current conclusion is that although the limits to prediction are real, it seems like AI could still substantially improve on human intelligence. The possibility of AI developing significant real-world capabilities in a relatively brief time seems like one that cannot be ruled out.

Before examining these questions, we need to consider the definition of "capability" in more detail, and justify our focus on intelligence as prediction ability.

Capability and intelligence as prediction ability

Bostrom (2014, p. 39) defines a superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest". Additionally, Bostrom

(2014, chap. 3) defines three subcategories of a superintelligence. A *speed superintelligence* thinks faster than humans; a *collective superintelligence* is composed of many smaller intellects whose overall performance outstrips that of existing cognitive systems; and a *quality superintelligence* is one that is at least as fast as a human mind, and vastly qualitatively smarter.

In a footnote to his original definition, Bostrom notes that this definition of superintelligence can be compared with Legg (2008), who defines intelligence as “an agent’s ability to achieve goals in a wide range of environments”.

This definition, originally from Legg & Hutter (2007a), draws on a collection of 70 definitions of intelligence (Legg & Hutter 2007b) from various professional groups, dictionaries, psychologists, and AI researchers. Legg & Hutter (2007a) argue that this definition summarizes the essential features in the various surveyed definitions, in that they generally discuss an individual who is interacting with some environment that is not fully known, trying to achieve various goals in that environment, and learning and exploring during that interaction.

Some definitions of intelligence list traits which are not explicitly included in this definition; for example, a group statement signed by 52 psychologists (Gottfredson 1997a) includes in intelligence "the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience". However, Legg & Hutter (2007a) argue that all of these abilities are ones that allow humans to achieve goals, so are implicitly included in the Legg & Hutter definition. Additionally, Legg & Hutter suggest that their definition is more general, as there could exist intelligences which did not have all of these specific capabilities, but did have alternative capabilities which allowed them to achieve their goals.

Legg & Hutter (2007a) offer a formalization of their definition, cast in a reinforcement learning framework. Briefly, the formalization involves an agent which interacts with an environment in discrete timesteps; on each timestep, the agent chooses an action and receives both an observation and a reward. An agent is (universally) intelligent to the extent that it can maximize its reward over the space of all environments drawn from a universal distribution.

This definition and formalization is a view of intelligent performance as a learning and prediction problem: an agent is intelligent to the extent that it can learn to predict, using the smallest possible set of observations, which of its actions will deliver the greatest amount of reward in the environment that it is interacting with.

Out of Bostrom’s (2014) superintelligence subtypes, a mind that was superintelligent under such a view would most likely fall under the category of a quality superintelligence. Some of the examples that Bostrom (2014) offers to illustrate the concept of quality intelligence include nonhuman animals that cannot achieve human cognitive capabilities even when “intensely trained by human instructors”, as well as human deficits such as autism spectrum disorders that may impair e.g. social functioning. Implicit in these examples is the notion that nonhuman

animals and individuals with cognitive deficits cannot achieve the same level of performance in various domains as unimpaired humans do, even when given the same opportunities to observe and learn about the domains in question. They lack the cognitive capabilities that would allow them to utilize their observations to learn to predict which kinds of actions would provide the greatest success in the relevant domains.

Under this view, we can more precisely rephrase our first question, “how much more capable can AIs become relative to humans”, as “how much better than humans can AIs become in *using small amounts of sense data to learn to predict which actions most effectively further their goals*”. For the purposes of this discussion, we will also assume that “predicting which actions most effectively further one’s goals” is an accurate characterization of what human expertise (in any given domain) means. As we will discuss in the following section, the foundation of human expertise lies in acquiring the necessary knowledge to instantly see, when faced with some situation, the right course of action for that situation.

The development of human expertise

Ideally, we might turn to theoretical AI research for a precise theory about acquiring cognitive capabilities. Unfortunately AI research is not at this point yet. Instead we will consider the research on human expertise and decision-making.

Expertise as mental representations

There exists a preliminary understanding, if not of the details of human decision-making, then at least the general outline. A picture that emerges from this research is that *expertise is about developing the correct mental representations* (Klein 1999; Ericsson & Pool, 2016).

A mental representation is a very general concept, roughly corresponding to any mental structure forming the content of something that the brain is thinking about (Ericsson & Pool, 2016).

Domain-specific mental representations are important because they allow experts to know what something means; know what to expect; know what good performance should feel like; know how to achieve the good performance; know the right goals for a given situation; know the steps necessary for achieving those goals; mentally simulate how something might happen; learn more detailed mental representations for improving their skills (Klein, 1999; Ericsson & Pool, 2016).

Although good decision-making is often thought of as a careful deliberation of all the possible options, such a type of thinking tends to be typical of novices (Klein, 1999). A novice will have to try to carefully reason their way through to an answer, and will often do poorly regardless,

because they do not know what things are relevant to take into account and which ones are not. An expert doesn't need to – they are experienced enough to instantly know what to do.

A specific model of expertise is the Recognition-Primed Decision-Making (RPD) model (Klein, 1999). First, a decision-maker sees some situation, such as a fire for a firefighter or a design problem for an architect. The situation may then be recognized as familiar, such as a typical garage fire. Recognizing a familiar situation means understanding what *goals* make sense and what should be focused on, which *cues* to pay attention to, what to *expect* next and when a violation of expectations shows that something is amiss, and knowing what the *typical ways of responding are.* Ideally, the expert will instantly know what to do.

The expectations arising from mental representations also give rise to *intuition*. As one example, Klein (1999) describes the case of a firefighter lieutenant responding to a kitchen fire in an ordinary one-story residential house. The lieutenant's crew sprayed water on the fire, but contrary to expectations, the water seemed to have little impact. Something about the situation seemed wrong to the lieutenant, who ordered his crew out of the house. As soon as they had left the house, the floor where they had been standing collapsed. If the firefighters had not pulled out, they would have fallen down to the fire raging in the basement. The lieutenant, not knowing what had caused him to give the order to withdraw, initially attributed the decision to some form of extra-sensory perception.

In a later interview, the lieutenant explained that he did not suspect that the building had a basement, nor that the seat of the fire was under the floor that he and his crew were standing on. However, several of his expectations of a typical kitchen fire were violated by the situation. The lieutenant was wondering why the fire did not react to water as expected, the room was much hotter than he would have expected out of a small kitchen fire, and while a heat that hot should have made a great deal of noise, it was very quiet. The mismatch between the expected pattern and the actual situation led to an intuitive feeling of not knowing what was going on, leading to the decision to regroup. This is intuition: an automatic comparison of the situation against existing mental representations of similar situations, guiding decision-making in ways whose reasons are not always consciously available.

In an unfamiliar situation, the expert may need to construct a *mental simulation* of what is going on, how things might have developed to this point, and what effect different actions would have. Had the floor mentioned in the previous example not collapsed, given time the firefighter lieutenant might have been able to put the pieces together and construct a narrative of a fire starting from the basement to explain the discrepancies. For a future-oriented example, a firefighter thinking about how to rescue someone from a difficult spot might mentally simulate where different rescue harnesses might be attached on the person, and whether that would exert dangerous amounts of force on them.

Mental representations are necessary for a good simulation, as they let the expert know what things to take into account, what things could plausibly be tried, and what effects they would

have. In the example, the firefighter's knowledge allows him to predict that specific ways of attaching the rescue harness would have dangerous consequences, while others are safe.

Developing mental representations

Mental representations are developed through practice. A novice will try out something and see what happens as a result. This gives them a rough mental representation and a prediction of what might happen if they try the same thing again, leading them to try out the same thing again or do something else instead.

Just practice isn't enough, however – there also needs to be feedback. Someone may do a practice drill over and over again and *think* that they are practicing and thus improving – but without some sign of how well that is going, they may just keep repeating the same mistakes over and over (Ericsson & Pool, 2016).

The importance of quality feedback is worth emphasizing. Skills do not develop unless there is feedback that is conducive to developing better mental representations. In fact, there are entire fields in which experienced practitioners are not much better than novices, because the field does not provide them with enough feedback. Shanteau (1992) provides the following breakdown of professions for which there is agreement on the nature of their performance:

Good performance	Bad performance
Weather Forecasters	Clinical Psychologists
Livestock Judges	Psychiatrists
Astronomers	Astrologers
Test Pilots	Student Admissions
Soil Judges	Court Judges
Chess Masters	Behavioral Researchers
Physicists	Counselors
Mathematicians	Personnel Selectors
Accountants	Parole Officers
Grain Inspectors	Polygraph (Lie Detector) Judges
Photo Interpreters	Intelligence Analysts
Insurance Analysts	Stock Brokers

In analyzing why some domains enable the development of genuine expertise and others don't, Shanteau identified a number of considerations that relate to the nature of feedback. In an occupation like weather forecasting, the criteria you use for forecasting are always the same; you will always be facing the same task and can practice it over and over; you get quick and feedback on whether your prediction was correct; you can use formal tools to analyze what you predicted would happen and why that prediction did or didn't happen; and things can be

analyzed in objective terms. This allows weather forecasters to develop powerful mental representations that get better and better at making the correct prediction.

Contrast this with someone like an intelligence analyst. The analyst may be called upon to analyze very different clues and situations; each of the tasks may be unique, making it harder to know which lessons from previous tasks apply; for many of the analyses, one might never know whether they were right or not; and questions about socio-cultural matters tend to be much more subjective than questions about weather, making objective analysis impossible. In short, for much of the work that the analyst does, there is simply no feedback available to tell whether the analyst has made the right judgment or not. And without feedback, there is no way to improve one's mental representations, and thus expertise.

A slightly different look on expertise is the heuristics & biases literature, which frequently portrays even experts as being easily mistaken. In contrast, the expertise literature that we have reviewed so far has viewed experts as being typically capable and as having trustworthy intuition. Kahneman & Klein (2009) make an attempt to reconcile the two fields, and come to agree that:

- Expert intuition may be trustworthy, if the intuition relates to a 'high-validity' domain and the expert has had a chance to learn the regularities in that domain.
- A domain is 'high-validity' if 'there are stable relationships between objectively identifiable cues and subsequent events or between cues and the outcomes of possible actions'.
- Medicine and firefighting have fairly high validity, whereas predictions of the future value of individual stocks and long-term[1] forecasts of political events are domains with practically zero validity.
- "Some [domains] are both highly valid and substantially uncertain. Poker and warfare are examples. The best moves in such situations reliably increase the potential for success."
- "[A domain] of high validity is a necessary condition for the development of skilled intuitions. Other necessary conditions include adequate opportunities for learning the [domain] (prolonged practice and feedback that is both rapid and unequivocal). If [a domain] provides valid cues and good feedback, skill and expert intuition will eventually develop in individuals of sufficient talent. "

This consensus is in line with what we have covered so far, though it also includes the consideration of validity. One cannot learn mental representations that would predict a domain or dictate the right actions for different situations in a domain, if that domain is simply too complicated or chaotic to be predicted. Kahneman & Klein (2009) provide an illustrative example of domain being simply too hard to interpret: the question of how the history of the 20th century would have been different if the fertilized eggs that became Hitler, Stalin and Mao had

[1] Kahneman & Klein do not define what they mean by 'long-term', but geopolitical events up to a year or so away can be predicted with reasonable accuracy, with the accuracy falling towards chance for events 3 to 5 years away. (Tetlock & Gardner 2015, p. 5).

been female. It seems clear that things would have developed very differently, but how exactly? There seems to be no way to know.

Meanwhile, practice does help in more predictable domains. A recent meta-analysis (Macnamara, Hambrick, & Oswald, 2014) on the effects of practice on skill found that the more predictable an activity was, the more practice contributed to performance in that activity.

Implications for AI

Having reviewed some necessary background, we will now finally get back to the topic of superintelligence capabilities.

Relevance for AI

Similarly to humans, AI systems cannot reach intelligent conclusions by a mere brute force calculation of every possibility. Rather, an intelligence needs to learn to exploit predictable regularities in the world in order to develop further. All machine learning based systems are based on this principle: they learn models of the world that are in this sense similar to the mental representations that humans learn.

However, the models employed by current machine learning systems are much more limited than the mental representations employed by humans (Lake et al. 2016). Machine learning systems are also developed for solving problems efficiently on existing computing hardware rather than for being biologically plausible. There is thus reason to expect even future AI systems to employ models which differ in various respects from the mental representations used by humans. As such, we will use the term “mental representations” when in the context of humans, and “models” when discussing the analogous structure in future AI systems.

In a sense, mental representations contain the optimal solutions to the problems at hand (Klein 1999): a human expert will have learned to identify the smallest set of cues that will let them know how to act in a certain situation; their mental representations encode information about how to choose the correct actions using the least amount of thought. In other words, an expert pays attention exactly to the features in the data which are relevant for making the decision, and acts accordingly. An AI’s models could use more data and become larger than human mental representations, and identify features which humans might have missed. There is however no advantage in using more data than necessary for making the correct decision, so at least a subset of the AI’s models is likely to be similar to mental representations in that they encode the smallest amount of features of the environment which allow for rapid and correct decision-making in a given context and for a given goal.

It is possible that AIs would *also* come to have models for which this characterization was a poor fit and which were tailored for taking better advantage of e.g. an AI’s ability to process

more data at a time. We will not examine this more speculative possibility, as for our argument it is unnecessary to consider hypothetical models which are *better* than human mental representations; we are focused on establishing the possibility that roughly human-like models would already be enough to enable superhuman capability[2].

Like with human experts, machine learning also tries to focus its analysis on exactly the right number of cues that will provide the right predictions, ignoring any irrelevant information. Traditional machine learning approaches have relied extensively on *feature engineering*, a labor-intensive process where humans determine which cues in the data are worth paying attention to.

A major reason behind the recent success of deep learning models is their capability for *feature learning* or *representation learning*: being able to independently discover high-level features in the data which are worth paying attention to, without (as much) external guidance (Bengio, Courville, & Vincent, 2012). Being able to identify and extract the most important features of the data allows the system to make its decisions based on the smallest amount of cues that allows it to reach the right judgment – just as human experts learn to identify the most relevant cues in the situations that they encounter.

Finally, the aspect of increasingly detailed mental representations giving an expert a yardstick to compare their performance against (Ericsson & Pool 2016) has an analogue in reinforcement learning methods. In deep reinforcement learning, a deep learning model learns to estimate how valuable a specific state of the world is, after which the system takes actions to move the world towards that state (Mnih et al., 2015). Similarly, a human expert comes to learn that specific states (e.g. a certain feeling in the body when diving) are valuable, and can then increasingly orient their behavior so as to achieve this state.

In summary, human experts use mental representations as the building blocks of their expertise, with the models employed by current state-of-the-art AI systems having a number of key similarities. As there have been no serious alternative accounts presented of how expertise might work,wewill assume that the capabilities of hypothetical superintelligences will depend, at least in part, on them developing the correct models to represent key features of the environment in a similar way as human mental representations do.

This paper set out to consider two main questions:

1. How much more capable can AIs become relative to humans?

[2] The reader may note that the AI possibly using many different kinds of models, some of them humanlike and some more advanced, has a parallel in the heterogeneity hypothesis of concepts (Machery 2009, 2010), according to which the mental representations of humans do not form a natural kind and actually consist of many different kinds of mental structures that are used in different situations and for different purposes.

2. How easily (in terms of time and resources required) could superhuman capability be acquired?

Let us now return to these.

The argument for AI's predictive capabilities being limited was that there are limits to prediction, and that predicting events an ever-increasing amount forward in time requires exponential reasoning power as well as measurement points, quickly becoming intractable. How capable could AI become despite these two points?

The components of human expertise might be roughly divided into two: building up a battery of accurate mental representations, and being able to use them for mental simulations. Similarly, approaches to artificial intelligence can roughly be divided into pattern recognition and model-building (Lake, Ullman, Tenenbaum, & Gershman, 2016), depending on whether patterns in data or models of the world are treated as the primary unit of thought.

As this kind of a distinction seems to emerge both from psychology and AI research,wewill assume that AI's expertise will also involve acquiring models (or equivalently, doing pattern recognition) as well as accurately using them in simulations. We will consider these two separately.

Simulation

Potential capability

An interesting look at the potential benefits offered by improved simulation ability come from looking at Philip Tetlock's Good Judgement Project (GJP), popularized in the book *Superforecasting* (Tetlock & Gardner, 2015)[3]. Participating in a contest to forecast the probability of various events, the best GJP participants – the so-called 'superforecasters' – managed to make predictions whose accuracy outperformed those of professional intelligence analysts working with access to classified data[4]. This is particularly interesting as the superforecasters had no particular domain expertise in answering most of the questions, with sample questions including ones such as

- Will North Korea launch a new multistage missile before May 10, 2014?

[3] Except for when citations to other content are explicitly included, all the discussion about superforecasters and the Good Judgment Project uses *Superforecasting* as its source.
[4] Though this claim needs to be treated with some caution, as no official information about the intelligence analysts' performance has been published. The claim is based on Washington Post editor David Ignatius writing that 'a participant in the project' had told him that superforecasters had 'performed about 30 percent better than the average for intelligence community analysts who could read intercepts and other secret data' (Ignatius, 2013). The intelligence community has neither confirmed nor denied this statement, and Philip Tetlock has stated that he believes it to be true.

- Will Russian armed forces enter Kharkiv, Ukraine, by May 10, 2014?
- Will there be a significant attack on Israeli territory before May 10, 2014?
- Will Robert Mugabe cease to be President of Zimbabwe by September 30, 2011?
- Will Greece remain a member of the EU through June 1, 2012?

Tetlock & Gardner report the superforecasters' accuracy in terms of Brier score, which is a scale between 0 and 2, with 0.5 indicating random guessing[5]. On this scale, superforecasters had a score of 0.25 at the end of GJP's first year, compared to 0.37 of the other forecasters participating in the project. By the end of the second year, superforecasters had improved their Brier score to 0.07 (Mellers et al., 2014). Superforecasters could also project further out in time: their accuracy at making predictions 300 days out was better as the other forecasters' accuracy at making predictions 100 days out. In terms of being on the right side of 50/50, GJP's best wisdom-of-the-crowd algorithms (deriving an overall prediction from the different forecasters' predictions) delivered a correct prediction on 86% of all daily forecasts (Tetlock, Mellers, & Rohrbaugh, 2014).

The superforecasters' success relied on a number of techniques, but a central one was the ability to consider and judge the relevance of a number of factors that might cause a prediction to become true or false. Tetlock & Gardner illustrate this technique by discussing how a superforecaster, Bill Flack, approached the question of whether an investigation of Yasser Arafat's remains would reveal traces of polonium, suggestive of Arafat having been poisoned by Israel.

Flack started by considering what it would take for the investigation to reach a particular outcome, and realized that he didn't know what the chances were of polonium traces surviving in a body for several years. He started by investigating how polonium testing worked, and concluded that enough polonium could in fact survive for it to be found in the testing.

Next, Flack considered what *could* cause polonium to end up in the body. Israel poisoning Arafat could have done it, but so could an Palestinian enemy that Arafat had. There was also the probability of the body being intentionally contaminated after Arafat's death, by some faction trying to frame Israel for the death. Each possibility made a positive test result more probable, based on how probable those individual possibilities were. Next Flack moved on to investigate what it would take for any of the possibilities to be true. For the case of Israel poisoning Arafat, it required Israel having access to polonium; Israel being willing to take the risk of intentionally poisoning him; and Israel having the means to poison Arafat with the polonium. These possibilities served as starting points for researching the probability of the "Israel poisoned Arafat" hypothesis, after which Flack would break down and investigate what it would take for the other hypotheses to be true.

[5] A version of the scale which ranges between 0 and 1 is also commonly used.

Tetlock does not go into detail about the prerequisites for being able to carry out such analysis – other than noting that it's slow and effortful – but there are some considerations that seem like plausible prerequisites. First, a person needs to have enough general knowledge to generate different possibilities for how an event could have come true. Next, they need the ability to analyze and investigate those possibilities further, either personally acquiring the relevant domain knowledge for evaluating their plausibility, or finding a relevant subject matter expert. In this example, Flack familiarized himself with the science of polonium testing until he was satisfied that it would be possible to detect polonium traces from a long time ago.

This suggests a general procedure which AI could also follow in order to predict the possibility of something in which it does not yet have expertise. An AI that was trying to predict the outcome of some specific question could work tap into its existing general knowledge in an attempt to identify relevant causal factors; if it failed to generate them, it could look into existing disciplines which seemed relevant for the question. For each identified possibility, it could branch off a new subprocess to do research into that particular direction, sharing information as necessary with a main process whose purpose was to integrate the insights derived from all the relevant searches.

Such a capability for several parallel streams of attention could provide a major advantage. A human researcher or forecaster who branches off to do research on a subquestion will need to make sure that they don't lose track of the big picture, and needs to have an idea of whether they are making meaningful progress on that subquestion and whether it would be better to devote attention to something else instead. To the extent that there can be several parallel streams of attention, these issues can be alleviated, with a main stream focusing on the overall question and substreams on specific subpossibilities.

How much could this improve on human forecasters? Forecasters performed better when they were placed on teams where they shared information between each other, which similarly allowed an extent of parallelism in prediction-making, in that different forecasters could pursue their own angles and directions in exploring the problem. The differences between individual forecasters and teams of forecasters with comparable levels of training ranged between 0.05 and 0.10 Brier points at the end of the first year, and between 0.02 and 0.08 Brier points at the end of the second year (Mellers et al., 2014). In humans however, it seems likely that the extent of parallelism was constrained by the fact that each forecaster had to independently familiarize themselves with much of the same material, and that their ability to share knowledge between each other was limited by the speed of writing and reading. This suggests a possibility for further improvement.

In general, accurate forecasting requires an ability to carry out sophisticated causal modeling about a variety of interacting factors. Tetlock & Gardner emphasize the extent to which superforecaster forums discuss many different "on the one hand"/"on the other hand" possibilities. In a discussion of whether Saudi Arabia might agree to OPEC production cuts in November 2014, one superforecaster noted that Saudi Arabia had large financial reserves so

could afford to let oil prices run low. On the other hand, he noted, Saudi Arabia needed to raise their social spending to bolster the support for the monarchy, but yet again, Saudi Arabian rulers might view the act of trying to control oil prices as futile. The superforecaster in question concluded that the question “felt no-ish, 80%”. (Saudis ended up not supporting production cuts.)

This suggests that AI with sufficient hardware capability could achieve considerable prediction ability by its capability to explore many different perspectives and causal factors at once. The simulations of humans tend to be limited to around three causal factors and six transition states (Klein, 1999). The discussion of the superforecasters clearly brought up many more possibilities, and their accuracy suggests moderate ability to integrate all those factors together. Yet comments such as 'feels no-ish' suggests that they still couldn’t construct a full-blown simulation in which the various causal factors would have influenced each other based on principled rules which could be inspected, evaluated, and revised based on feedback and accuracy. This seems especially plausible given that Klein speculates the limits in the size of human simulations to come from working memory limitations.

AI systems with larger working memory capacities might be able to construct much more detailed simulations. Contemporary computer models can involve simulations with thousands or tens of thousands variables, though flexibly incorporating diverse models into a single simulation will probably take considerably more memory and computing power than what is used in today’s models.

Example: parallel streams of attention with a LIDA-like architecture

How could different streams of attention within AI share information between each other? Recall that we have defined the development of expertise as the ability to accumulate patterns which are used to identify relevant cues and to indicate what predictions should be derived out of those. A computational model for attention and consciousness is Global Workspace Theory (Baars, 2002; 2005), of which a particular AI implementation is the LIDA model (Franklin & Patterson, 2006; Franklin, Madl, D’Mello, & Snaider, 2014; Madl, Franklin, Chen, Montaldi, & Trappl, 2016). LIDA is a model of the mind that is inspired by psychological and neuroscientific research and attempts to capture its main mechanisms.

We can use LIDA to get a rough example of what having several 'streams of attention' would mean, and how information could be exchanged between them. The purpose of this example is not to suggest that an AI would necessarily work by this mechanism, but merely to make the speculation about streams of attention slightly more grounded in existing theories of how a general intelligence (the human mind) might work. Thus, to the extent that LIDA is correct as a model of human intelligence, and to the extent that the example in this box is correct about LIDA allowing for there to be several attentional streams at the same time, this provides some

information about it being possible to have several such streams in minds in general, and how that might concretely work.

LIDA works by means of an understand-attend-act cycle. In each cycle, low-level sensory information is initially interpreted so as to associate it with higher-level concepts to form a 'percept', which is then sent to a workspace. In the workspace, the percept activates further associations in other memory systems, which are combined with the percept to create a Current Situational Model, an understanding of what is going on at this moment.

The entirety of the Current Situational Model is likely to be too complex for the agent to process, so it needs to select a part of it to elevate to the level of conscious attention to be acted upon. This is carried out using 'attention codelets', small pieces of code that attempt to train attention on some particular piece of information, each with their own set of concerns of what is important. Attention codelets with matching concerns form coalitions of what to attend, competing against other coalitions. Whichever coalition ends up winning the competition will have its chosen part of the Current Situational Model 'become conscious', broadcast to the rest of the system, and particularly Procedural Memory.

The Procedural Memory holds schemes, or templates of different actions that can be taken in different contexts. Schemes which include a context or an action that matches the contents of the conscious broadcast become available as candidates for possible actions. They are copied to the Action Selection mechanism, which chooses a single action to perform. The selected action is further sent to Sensory-Motor Memory, which contains information of how exactly to perform the action. The outcome of taking this action manifests itself as new sensory information, beginning the cognitive cycle anew.

Here is a description of how this process – or something like it – might be applied in the case of AI seeking to predict the outcome of a specific question, such as the 'will Saudi Arabia agree to oil production cuts' question discussed above. The decision to consider this question has been made in an earlier cognitive cycle, and information relevant to it is now available in the inner environment and the Current Situational Model. The concepts of Saudi Arabia and oil production trigger several associations in the AI's memory systems, such as the fact that oil prices will affect Saudi Arabia's financial situation, and that oil prices are also influenced by other factors such as global demand. Two coalitions of attention codelets might form, one focusing on the current financial situation and another on influences on oil prices.

In LIDA, these codelets would normally compete, and one of them would win and trigger a specific action, such as a deeper investigation of Saudi Arabia's financial situation. In our hypothetical AI however, it might be enough that both coalitions manage to exceed some threshold level of success, indicating them both to be potentially relevant. In that case, new instances of the Procedural Memory, Action Selection and Sensory-Motor Memory mechanisms might be initialized, with one coalition sending its contents to the first set of

instances and the other to another. These streams could then independently carry out searches of the information that was deemed relevant, also having their own local Situational Models and Workspaces focusing on content relevant for this search. As they worked, these streams would update the various memory subsystems with the results of their learning, making new associations and attention codelets available to all attentional streams. Their functioning could be supervised by a general high-level attention stream, whose task was to evaluate the performance of the various lower-level streams and allocate resources between them accordingly.

These simulations do not necessarily need to incorporate an exponentially increasing number of variables in order to achieve better prediction accuracy. As previously noted, superforecasters were more accurate at making predictions 300 days out than the rest of the forecasters in GJP were at making predictions 100 days out. Given that at least some of the superforecasters only used a few hours a day on making their predictions, and that they had many predictions to rate, they probably did not consider a *vastly* larger amount of factors than the rest of the forecasters.

Klein (1999) offers an example of a professor who used three causal factors (the rate of inflation, the rate of unemployment, and the rate of foreign exchange) and a few transitions to relatively accurately simulate how the Polish economy would develop in response to the decision to convert from socialism to a market economy. In contrast, less sophisticated experts could only name two variables (inflation and unemployment) and not develop any simulations at all, basing their predictions mostly on their ideological leanings.

Having large explicit models also allows for the models to be adjusted in response to feedback. The professor's estimate was in many extents correct, but failed to predict the government being less ruthless and more cautious than it had said it would be closing down unproductive plants. The government's caution could thus be added as an additional variable to be considered for the next model. The addition of this variable alone might then considerably increase the accuracy of the simulation.

Tetlock & Gardner report that the superforecasters used highly granular probability estimates – carefully thinking about whether the probability of an event was 3% as opposed to 4%, for instance – and that the granularity actually contributed to accuracy, with the predictions getting less accurate if they were rounded to the closest 5%. Given that such granularity was achieved by integrating various possibilities and considerations, it seems like an ability to consider and integrate an even larger amount of possibilities might provide even increased granularity, and thus a prediction edge.

In summary, AI could be able to run vastly larger simulations than humans could, with this possibility being subject to computing power limitations; given this, its simulations could also be explicit, allowing it to adjust and correct them in response to feedback to provide improved prediction accuracy; and it could have several streams of attention running concurrently and

sharing information between each other. Existing evidence from human experts suggests that large increases to prediction capability might not necessarily need a large increase in the number of variables considered, and that even small increases can provide considerable additional gains.

The amount of predictive edge that this could give to an AI as compared to a human or a group of humans is unclear, but humans do tend to prefer simple stories and explanations that are compact enough that all of the important details can be kept in mind at once. Simple hypotheses often turn out to be insufficient because the world is more complicated than a simple hypothesis allows for. Even in domains such as engineering, where there exist formal ways of modeling the entire domain, a task such as the design of a modern airplane or operating system contains too much complexity for a single person to comprehend. While the impact of uncertainty can never be eliminated, being able to take more of the world's underlying complexity into account than humans do, may provide an AI with a predictive edge at least in some domains.

Rate of capability growth

How fast could AI develop the ability to run comprehensive and large simulations?[6] Creating larger simulations than humans have access to seems to require extensive computational resources, either from hardware or optimized software. As an additional consideration, we have previously mentioned limited working memory restricting the capabilities of humans, but human working memory is *not* the same thing as RAM in computer systems. If one were running a simulation of the human brain in a computer, one could not increase the brain's available working memory simply by increasing the amount of RAM the simulation had access to. Rather, it has been hypothesized that working memory differences between individuals may reflect things such as the ability to discriminate between relevant and irrelevant information (Unsworth & Engle, 2007), which could be related to things like brain network structure and thus be more of a software than a hardware issue.[7] Yudkowsky (2013) notes that if increased intelligence would be a simple matter of scaling up the brain, the road from chimpanzees to humans would likely have been much shorter, as simple factors such as brain size can respond rapidly to evolutionary selection pressure.

Thus, advances in simulation size depend on progress in both hardware and algorithms. Hardware progress in hard to predict, but advances in algorithmic capabilities seem doable using mostly theoretical and mathematical research. This would require the development of expertise in mathematics, programming, and theoretical computer science.

[6] This section does not consider how fast the AI could develop the necessary mental representations to be used in the simulations. That question will be discussed in the next section.

[7] Though it is worth noting that *g* does correlate to some extent with brain size, with a mean correlation of 0.4 in measurements that are obtained using brain imaging as opposed to external measurements of brain size (Rushton & Ankney, 2009). This would seem to suggest that the raw number of neurons and thus 'general hardware capacity' would also be relevant.

Much of mathematical problem-solving is about having a library of procedures, reformulations, and heuristics that one can try (Polya, 1990), as well as developing a familiarity and understanding of many kinds of mathematical results, which one may then later on recognize as relevant. This seems like the kind of task that relies strongly on pattern-matching abilities, and might in principle be in reach by an advanced deep reinforcement learning system that was fed a sufficiently large library of heuristics and worked proofs to let it develop superhuman mathematical intuition[8]. Modern-day theorem provers often know what kinds of steps are valid, but not which steps are worth taking; merging them with the 'artificial intuition' of deep reinforcement learning systems might eventually produce systems with superhuman mathematical ability.

Progress in this field could allow AI systems to achieve superhuman abilities in math research, considerably increasing their ability to develop more optimized software to take full advantage of the available hardware. To the extent that relatively small increases in the number of variables considered in a high-level simulation would allow for dramatically increased prediction ability (as is suggested by e.g. the superforecasters being better predictors with thrice the prediction horizon of less accurate forecasters), moderate increases in the size of the AI's simulations could translate to drastic increases in terms of real-world capability.

Yudkowsky (2013) notes that although the evolutionary record strongly suggests that algorithmic improvements were needed for taking us from chimpanzees to humans, the record rules out exponentially increasing hardware always being needed for linear cognitive gains: the size of the human brain is only four times that of the chimpanzee brain. This further suggests that relatively limited improvements could allow for drastic increases in intelligence.

Pattern recognition

The capability to run large simulations isn't enough by itself. The AI also needs to acquire a sufficiently large number of patterns to be included in the simulations, to predict how different pieces in the simulation behave.

Potential capability

When it comes to well-defined tasks, current AI systems excel at pattern recognition, being able to analyze vast amounts of data and build them into an overall model, finding regularities that human experts never would have. For instance, human experts would likely have been unable to anticipate that men who 'like' the Facebook page 'Being Confused After Waking Up From Naps' are more likely to be heterosexual (Kosinski, Stillwell, & Graepel, 2013). Similarly, the Go-playing AI AlphaGo, whose good performance against the expert player Lee Sedol could to a large extent be attributed to its built-up understanding of the kinds of board patterns that

[8] See Whalen (2016) for preliminary work in this direction.

predict victory, managed to make moves that Go professionals watching the game considered creative and novel.

The ability to find subtle patterns in data suggests that AI systems might be able to make predictions in domains which humans currently consider impossible to predict. We previously discussed the issue of the (predictive) *validity* of a domain, with domains being said to have higher validity if 'there are stable relationships between objectively identifiable cues and subsequent events or between cues and the outcomes of possible actions' (Kahneman & Klein, 2009). A field could also be valid despite being substantially uncertain, with warfare and poker being listed as examples of fields that were valid (letting a skilled actor improve their average performance) despite also being highly uncertain (with good performance not being guaranteed even for a skilled actor).

We already know that the validity of a field also depends on an actor's cognitive and technological abilities. For example, weather forecasting used to be a field in which almost no objectively identifiable cues were available, relying mostly on guesswork and intuition, but the development of modern meteorological theory made it a much more valid field (Shanteau, 1992). Thus, even fields which have low validity to humans with modern-day capabilities, could become more valid for more advanced actors.

A possible example of a domain that is currently relatively low-validity, but which could become substantially more valid, is that of predicting the behavior of individual humans. Machine learning tools can already generate personality profiles harvested from people's Facebook 'likes' that are slightly more accurate than the profiles made by people's human friends (Youyou et al. 2015), and can be used to predict private traits such as sexual orientation (Kosinski et al. 2013). This has been achieved using a relatively limited amount of data and not much intelligence; a more sophisticated modeling process could probably make even better predictions from the same data.

Taleb (2007) has argued for history being strongly driven by 'black swan' events, events with such a low probability that they are unanticipated and unprepared for, but which have an enormous impact on the world. To the extent that this is accurate, it suggests limits on the validity of prediction. However, Tetlock & Gardner (2015) argue that while the black swans themselves may be unanticipated, once the event has happened its consequences may be much easier to predict. Although superforecasters have shown no ability to predict black swans such as the 9/11 terrorist attacks, they could predict the answers to questions like "Will the United States threaten military action if the Taliban don't hand over Osama bin Laden?" and "Will the Taliban comply?".

Thus, even though AI might be unable to predict some very rare events, once those events have happened, it could utilize its built-up knowledge of how people typically react to different events in order to predict the consequences better than anyone else.

Rates of capability growth

How quickly could AI acquire more detailed models? Here again opinions differ. Hibbard (2016) argues, based on Mahoney's (2008) argument for intelligence being a function of both resources and knowledge, that explosive growth is unlikely. Benthall (2017) makes a similar argument. On the other hand, authors such as Bostrom (2014) and Yudkowsky (2008) suggest the possibility for fast increases.

How to improve learning speed?

We know that among humans, there are considerable differences in the extent to which people learn. Human cognitive differences have a strong neural and genetic basis (Deary, Penke, & Johnson, 2010), and strongly predict academic performance (Deary et al., 2007), socio-economic outcomes (Strenze, 2007), and job performance and the effectiveness of on-the-job learning and experience (Gottfredson, 1997b). There also exist child prodigies who before adolescence achieve a level of performance comparable to an adult professional, without having been able to spend comparable amounts of time training (Ruthsatz, Ruthsatz, & Stephens, 2013). In general, some people are able to learn faster from the same experiences, notice relevant patterns faster, and continue learning from experience even past the point where others cease to achieve additional gains.[9]

While there is so far no clear consensus on why some people learn faster than others, there are some clear clues. Individual differences in cognitive abilities may be a result of differences in a

---

[9] Readers who are familiar with the 'deliberate practice' literature may wonder if that literature might not contradict these claims about the impact of intelligence. After all, the deliberate practice research suggests that talent is irrelevant, and that deliberate, well-supervised training is the only thing that matters.

However, as noted by the field's inventor, deliberate practice is a concept that is applicable to some very specific – one might even say artificial – domains. Deliberate practice can only be applied in fields in which there are objective metrics, highly developed objectively-measurable expertise, and active competition to improve the existing practices. Areas that don't qualify are "*anything in which there is little or no direct competition, such as gardening and other hobbies, for instance, and many of the jobs in today's workplace— business manager, teacher, electrician, engineer, consultant, and so on*", as there are no objective criteria for performance (Ericsson & Pool 2016).

Fields that have well-defined, objective criteria for good performance are ones which are the easiest to master using even current-day AI methods – in fact, they're basically the only ones that can be truly mastered using current-day AI methods.

A somewhat cheeky way to summarize these results would be by saying that, in the kinds of fields that could be mastered by AI methods that exhibit no general intelligence, general intelligence isn't the most important thing. This even seems to be Ericsson's own theoretical stance: that in these fields, general intelligence eventually ceases to matter because the expert will have developed specialized mental representations that they can just rely on in every situation. So these results are not very interesting to those of us who are interested in domains that *do* require general intelligence.

combination of factors, such as working memory capacity, attention control, and long-term memory (Unsworth et al., 2014). Ruthsatz et al. (2013), in turn, note that '*child prodigies' skills are highly dependent on a few features of their cognitive profiles, including elevated general IQs, exceptional working memories, and elevated attention to detail*'.

Many tasks require paying attention to many things at once, with a risk of overloading the learner's working memory before some of the performance has been automated. For an example, McPherson & Renwick (2001) consider children who are learning to play instruments, and note that children who had previously learned to play another instrument were faster learners. They suggest this to be in part because the act of reading musical notation had become automated for these children, saving them from the need to process notation in working memory and allowing them to focus entirely on learning the actual instrument.

This general phenomenon has been recognized in education research. Complex activities that require multiple subskills can be hard to master even if the students have moderate competence in each individual subskill, as using several of them at the same time can produce an overwhelming cognitive load (Ambrose et al. 2010, chap. 4). Recommended strategies for dealing with this include reducing the scope of the problem at first and then building up to increasingly complex scopes. For instance, '*a piano teacher might ask students to practice only the right hand part of a piece, and then only the left hand part, before combining them*' (ibid).

An increased working memory capacity, which is empirically associated with faster learning capabilities, could theoretically assist in learning in allowing more things to be comprehended simultaneously without them overwhelming the learner. Thus, AI with a large working memory could learn and master at once much more complicated wholes than humans.

Additionally, we have seen that a key part of efficient learning is the ability to monitor one's own performance and to notice errors which need correcting; this seems in line with cognitive abilities correlating with attentional control and elevated attention to detail. McPherson & Renwick (2001) also remark on the ability of some students to play through a piece with considerably fewer errors on their second run-through than the first one, suggesting that this indicates '*an outstanding ability to retain a mental representation of [...] performance between run-throughs, and to use this as a basis for learning from [...] errors*'. In contrast, children who learned more slowly seemed to either not notice their mistakes, or alternatively to not remember them when they played the piece again.

Whatever the AI analogues of working and long-term memory, attentional control, and attention to detail are, it seems at least plausible that these could be improved upon by drawing exclusively on relatively theoretical research and in-house experiments. This might enable AI to both absorb vast datasets, as current-day deep learning systems do, *and* also learn from superhumanly small amounts of data.

Limits of learning speed

How much can the human learning speed be improved upon? This remains an open question. There are likely to be sharply diminishing returns at some point, but we do not know whether they are near the human level. Human intelligence seems constrained by a number of biological and physical factors that are unrelated to gains from intelligence. Plausible constraints include the size of the birth canal limiting the volume of human brains, the brain's extensive energy requirements limiting the overall amount of cells, limits to the speed of signaling in neurons, an increasing proportion of the brain's volume being spent on wiring and connections (rather than actual computation) as the number of neurons grows, and inherent unreliabilities in the operation of ion channels (Fox, 2011). There doesn't seem to be any obvious reason for why the threshold for diminishing gains from intelligence to learning speed would just happen to coincide with the level of intelligence allowed by our current biology. Alternatively, there could have been diminishing returns all along, but ones which still made it worthwhile for evolution to keep investing in additional intelligence.

The available evidence also seems to suggest that within the human range at least, increased intelligence continues to contribute to additional gains. The Study of Mathematically Precocious Youth (SMPY) is a 50-year longitudinal study involving over 5,000 exceptionally talented individuals identified between 1972 and 1997. Despite its name, many its participants are more verbally than mathematically talented. The study has led to several publications; among others, Wai et al. (2005) and Lubinski & Benbow (2006) examine the question of whether ability differences within the top 1% of the human population make a difference in life.

Comparing the top (Q4) and bottom (Q1) quartiles of two cohorts within this study shows both to significantly differ from the ordinary population, as well as from each other. Out of the general population, about 1% will obtain a doctoral degree, whereas 20% of Q1 and 32% of Q4 did. 0.4% of Q1 achieved tenure at a top-50 US university, as did 3% of Q4. Looking at a 1 to 10,000 cohort, 19% had earned patents, as compared to 7.5% of the Q4 group, 3.8% of the Q1 group, or 1% of the general population.

It is important to emphasize that the evidence we've reviewed so far does not merely mean that AI could potentially learn faster in terms of time: it also suggests that the AI could potentially learn faster *in terms of training data*. The smaller datasets AI needs in order to develop accurate models, the faster it can adapt to new situations.

Besides the considerations we have already discussed, there seems to be potential for accelerated learning through more detailed analysis of experiences. For example, chess players improve most effectively by studying the games of grandmasters, and trying to predict what moves the grandmasters would have made in any situation. When the grandmaster play deviates from the move that the student would have made, the student goes back to try to see what they missed (Ericsson & Pool, 2016). This kind of detailed study is effortful however, and can only be sustained for limited amounts at a time. With enough computational resources, the

AI could routinely run this kind of analysis on all sense data it received, constantly attempting to build increasingly detailed models that would correctly predict the data.

How much interaction is needed?

Some commentators, such as Hibbard (2016) argue that knowledge requires interaction with the world, so the AI would be forced to learn over an extended period of time as the interaction takes time.

From our previous review, we know that feedback is needed for the development of expertise. However, one may also get feedback from studying static materials. As we noted before, chess players spend more time studying published matches and trying to predict the grandmaster moves – and then getting feedback when they look up the next move and have their prediction confirmed or falsified – than they do actually playing matches against live opponents (Ericsson & Pool, 2016). The Go-playing AlphaGo system did not achieve its skill by spending large amounts of time playing human opponents, but rather studying the games of humans and playing games against itself (Silver et al. 2016). And while any individual human can only study a single game at a time, AI systems could study a vast number of games in parallel and learn from all of them[10].

An important difference is that domains such as chess and Go are formally specified domains, which AI can perfectly simulate. For a domain such as social interaction, the AI's ability to accurately simulate the behavior of humans is limited by its current competence in the domain. While it can run a simulation based on its existing model of human behavior, predicting how humans would behave based on that model, it needs external data in order to find out how accurate its prediction was.

This is not necessarily a problem however, given the vast (and ever-increasing) amount of recorded social interaction happening online. YouTube, e-mail lists, forums, blogs, and social media services all provide rich records of various kinds of social interaction, for AI to test its predictive models against without needing to engage in interaction of its own. Scientific papers – increasingly available on an open access basis – on topics such as psychology and sociology offer additional information for the AI to supplement its understanding with, as do various guides to social skills. All of this information could be acquired simply by downloading it, with the main constraints being the time needed to find, download, and process the data, rather than time needed for social interactions.

As noted earlier, relatively crude statistical methods can already extract relatively accurate psychological profiles out of data such as people's Facebook 'likes' (Kosinski et al., 2013, Youyou et al., 2015), giving reason to suspect that a general AI could develop very accurate predictive abilities given the kind of a process described above.

[10] See Mnih et al. (2016) for a discussion of how incorporating parallel learning improves upon on modern deep learning systems.

Several other domains, such as software security and mathematics seem similarly amenable to being mastered largely without needing to interact with the world outside the AI, other than searching for relevant materials. Some domains such as physics would probably need novel experiments, but AI focusing on the domains that were the easiest and fastest for it to master might find sufficient sources of capability from those alone.

Given the above considerations, it does not seem like AI's speed of learning would necessarily be strongly interaction-constrained.

Conclusions

We set out to consider the fundamental practical limits of intelligence, and the limits to how quickly an AI system could acquire very high levels of capability.

Fictional representations of high intelligence often depict a picture of geniuses as masterminds who have an almost godlike prediction ability, laying out intricate multi-step plans where every contingency is planned for in advance (TVTropes 2017a). When discussing "superintelligent" AI systems, one might easily think that the discussion was postulating something along the lines of those fictional examples, and rightly reject it as unrealistic.

Given what we know about the limits of prediction, for AI to make a single plan which takes into account every possibility is surely impossible. However, having reviewed the science of human expertise, we have found that experts who are good at their domains tend to develop powerful mental representations which let them react to various situations as they arise, and to simulate different plans and outcomes in their heads.

Looking from humans to AIs, we have found that AI might be able to run much more sophisticated mental simulations than humans could. Given human intelligence differences and empirical and theoretical considerations about working memory being a major constraint for intelligence, the empirical finding that increased intelligence continues to benefit people throughout the whole human range, and the observation that it would be unlikely for the theoretical limits of intelligence to coincide with the biological and physical constraints that human intelligence currently faces, it seems like AIs could come to learn considerably faster from data than humans do. It also seems like in many domains, this could be achieved by using existing materials as a source of feedback for predictions, without necessarily being constrained by time taken for interacting with the external world.

Thus, it looks that even though an AI system couldn't make a single superplan for world conquest right from the beginning, it could still have a superhuman ability to adapt and learn from changing and novel situations, and react to those faster than its human adversaries. As an analogy, experts playing most games can't precompute a winning strategy right from the first

move either, but they can still react and adapt to the game's evolving situation better than a novice can, enabling them to win[11].

Many of the hypothetical advantages – such as a larger working memory, the ability to consider more possibilities at once, and the ability to practice on many training instances in parallel – that AI might have seem to depend on available computing power. Thus the amount of hardware the AI had at its disposal could limit its capabilities, but there exists the possibility of developing better-optimized algorithms by initially specializing in fields such as programming and theoretical computer science, which the AI might become very good at.

One consideration which we have not yet properly addressed is the technology landscape at the time when the AI arrives (Tomasik 2014/2016, sec. 7). If a general AI can be developed, then various forms of sophisticated narrow AI will also be in existence. Some of them could be used to detect and react to a general AI, and tools such as sophisticated personal profiling for purposes of social manipulation will likely already be in existence. Considering how these influence the considerations discussed here is an important question, but one which is outside the scope of this article.

In summary, even if AI could not create a complete master plan from scratch, there seems to be a reasonable chance that could still come to substantially outperform humans in many domains, developing and using superior expertise than what humans were capable of. How fast AI systems could develop to such a level would depend on the speed at which algorithmic and hardware improvements became available. They could potentially be very fast, if e.g. the required algorithmic insights were more on the level of scaling up the size of the AI's simulations and number of attentional streams, rather than requiring any genuinely new ideas compared to what allowed the AI to achieve a rough human level in the first place.

Acknowledgments

Thank you to David Althaus, Stuart Armstrong, gwern branwen, Bill Hibbard, David Krueger, Josh Marlow, Carl Shulman, Brian Tomasik, and two anonymous reviewers on helpful comments on this paper.

References

Anderson, M. (2010). Problem Solved: Unfriendly AI. Retrieved September 27, 2016, from

[11] This is to say, while we concluded that the fictional trope of a "Xanatos Gambit" (TVTropes 2017a) is unrealistic, a much more accurate description of how a superintelligent AI actually acted could be the one of "Xanatos Speed Chess", in which complex plans are constantly revised as the situation progresses (TVTropes 2017b).

http://hplusmagazine.com/2010/12/15/problem-solved-unfriendly-ai/

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv:1606.06565 [cs.AI]*

Baars, B. J. (2002). The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences*, *6*(1), 47–52.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

Bengio, Y., Courville, A., & Vincent, P. (2012). *Representation Learning: A Review and New Perspectives*. *arXiv:1206.5538 [cs.LG]*

Benthall, S. (2017). Don't Fear the Reaper: Refuting Bostrom's Superintelligence Argument. *arXiv [cs.AI]* Retrieved from https://arxiv.org/abs/1702.08495

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

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 553-570). Springer International Publishing.

Buizza, R. (2002). *Chaos and weather prediction*. ECMWF. Retrieved from http://www.ecmwf.int/sites/default/files/Chaos%20and%20weather%20prediction.pdf

Deary, I. J., Penke, L., & Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews. Neuroscience*, *11*(3), 201–211. https://doi.org/10.1038/nrn2793

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13. https://doi.org/10.1016/j.intell.2006.02.001

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the New Science of Expertise*. Houghton Mifflin Harcourt.

Fox, D. (2011). The Limits of Intelligence. *Scientific American*. Retrieved from http://www.cs.virginia.edu/~robins/The_Limits_of_Intelligence.pdf

Franklin, S., Madl, T., D'Mello, S., & Snaider, J. (2014). LIDA: A Systems-level Architecture for Cognition, Emotion, and Learning. *IEEE Transactions on Autonomous Mental Development*, *6*(1). https://doi.org/10.1109/TAMD.2013.2277589

Franklin, S., & Patterson, F. G., Jr. (2006). The LIDA architecture: Adding new modes of learning to an intelligent, autonomous software agent. Presented at the Integrated Design and Process Technology, IDPT-2006. Retrieved from

http://ccrg.cs.memphis.edu/assets/papers/zo-1010-lida-060403.pdf

Future of Life Institute (2015). An Open Letter: Research Priorities for Robust and Beneficial Artificial Intelligence. Retrieved from https://futureoflife.org/ai-open-letter/ .

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24(1),* 13-23.

Gottfredson, L. S. (1997b). Why g matters: The complexity of everyday life. *Intelligence*, *24*(1), 79–132. https://doi.org/10.1016/S0160-2896(97)90014-3

Hibbard, B. (2016). A Defense of Humans for Transparency in Artificial Intelligence. Retrieved September 28, 2016, from http://www.ssec.wisc.edu/~billh/g/transparency_defense.html

Ignatius, D. (2013). David Ignatius: More chatter than needed. *The Washington Post*. Retrieved from https://www.washingtonpost.com/opinions/david-ignatius-more-chatter-than-needed/2013/11/01/1194a984-425a-11e3-a624-41d661b0bb78_story.html

Kahneman, D., & Klein, G. (2009). Conditions for Intuitive Expertise. A Failure to Disagree. *The American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Klein, G. (1999). *Sources of Power: How People Make Decisions*. MIT Press. Retrieved from https://books.google.de/books?id=nn1kGwL4hRgC

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). *Building Machines That Learn and Think Like People*. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/1604.00289

Lawrence, N. (2016). Future of AI 6. Discussion of 'Superintelligence: Paths, Dangers, Strategies.' Retrieved September 27, 2016, from http://inverseprobability.com/2016/05/09/machine-learning-futures-6

Legg, S. (2008). *Machine super intelligence* (Doctoral dissertation, Università della Svizzera italiana).

Legg, S., & Hutter, M. (2007a). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391-444.

Legg, S., & Hutter, M. (2007b). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and applications*, *157*, 17-24.

Lubinski, D., & Benbow, C. P. (2006). Study of Mathematically Precocious Youth After 35 Years: Uncovering Antecedents for the Development of Math-Science Expertise. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Machery, E. (2009). Doing without concepts. Oxford University Press.

Machery, E. (2010). Précis of doing without concepts. *Behavioral and Brain Sciences, 33(2-3)*, 195-206.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: a meta-analysis. *Psychological Science*, *25*(8), 1608–1618. https://doi.org/10.1177/0956797614535810

Madl, T., Franklin, S., Chen, K., Montaldi, D., & Trappl, R. (2016). Towards real-world capable spatial memory in the LIDA cognitive architecture. *Biologically Inspired Cognitive Architectures*, *16*, 87–104. https://doi.org/10.1016/j.bica.2016.02.001

Mahoney, M. (2008). *A Model for Recursively Self Improving Programs*. Retrieved from http://mattmahoney.net/rsi.pdf

Martela, F. (2016) Törmääkö tekoäly älykkyyden ylärajaan? *Tivi.fi.* Retrieved from http://www.tivi.fi/blogit/tormaako-tekoaly-alykkyyden-ylarajaan-6584349 .

McPherson, G. E., & Renwick, J. M. (2001). A Longitudinal Study of Self-regulation in Children's Musical Practice. *Music Education Research*, *3*(2), 169–186. https://doi.org/10.1080/14613800120089232

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., ... & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*. Retrieved from http://www.jmlr.org/proceedings/papers/v48/mniha16.pdf

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Polya, G. (1990). *How to Solve It: A New Aspect of Mathematical Method* (New edition). Penguin Books, Limited (UK). Retrieved from https://www.amazon.com/How-Solve-Mathematical-Penguin-Science/dp/0140124993

Rushton, J. P., & Ankney, C. D. (2009). Whole brain size and general mental ability: a review.

The International Journal of Neuroscience, *119*(5), 691–731. https://doi.org/10.1080/00207450802325843

Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine, 36*(4), 105-114.

Ruthsatz, J., Ruthsatz, K., & Stephens, K. R. (2013). Putting practice into perspective: Child prodigies as evidence of innate talent. *Intelligence*, *45*, 60–65. https://doi.org/10.1016/j.intell.2013.08.003

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes*, (53), 252–266. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.625.1063&rep=rep1&type=pdf

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529(7587)*, 484-489.

Soares, N., & Fallenstein, B. (2014). Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/files/TechnicalAgenda.pdf .

Sotala, K., & Yampolskiy, R. V. (2015). Responses to catastrophic AGI risk: a survey. *Physica Scripta*, *90*(1), 018001. https://doi.org/10.1088/0031-8949/90/1/018001

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence*, *35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004

Susan A. Ambrose, Michael W. Bridges, Michele DiPietro, Marsha C. Lovett, Marie K. Norman, Richard E. Mayer. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Jossey-Bass. Retrieved from http://teaching.temple.edu/sites/tlc/files/resource/pdf/What%20Factors%20Motivate%20Students%20to%20Learn_.pdf

Taleb, N. N. (2007). Black Swans and the Domains of Statistics. *The American Statistician*, *61*(3), 198–200. https://doi.org/10.1198/000313007X219996

Taylor, J., Yudkowsky, E., LaVictoire, P. & Critch, A. (2016) Alignment for advanced machine learning systems. *Machine Intelligence Research Institute*. Retrieved from https://intelligence.org/files/AlignmentMachineLearning.pdf .

Tetlock, P. E., Mellers, B. A., & Rohrbaugh, N. (2014). Forecasting Tournaments Tools for Increasing Transparency and Improving the Quality of Debate. *Current Directions in*. Retrieved from http://cdp.sagepub.com/content/23/4/290.short

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown.

Tomasik, B. (2014). Artificial Intelligence and Its Implications for Future Suffering. Retrieved September 28, 2016, from https://foundational-research.org/artificial-intelligence-and-its-implications-for-future-suffering#reply-to-bostroms-arguments-for-a-hard-takeoff

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

TVTropes (2017a). Xanatos Gambit. http://tvtropes.org/pmwiki/pmwiki.php/Main/XanatosGambit

TVTropes (2017b). Xanatos Speed Chess. http://tvtropes.org/pmwiki/pmwiki.php/Main/XanatosSpeedChess

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(1), 104–132. https://doi.org/10.1037/0033-295X.114.1.104

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. https://doi.org/10.1016/j.cogpsych.2014.01.003

Wai, J., Lubinski, D., & Benbow, C.P. (2005) Creativity and Occupational Accomplishments Among Intellectually Precocious Youths: An Age 13 to Age 33 Longitudinal Study. *Journal of Educational Psychology, 97(3),* 484-492.

Whalen, D. (2016). *Holophrasm: a neural Automated Theorem Prover for higher-order logic*. *arXiv [cs.AI]*. Retrieved from http://arxiv.org/abs/1608.02644

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(4), 1036–1040. https://doi.org/10.1073/pnas.1418680112

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In M. M. Ć. Nick Bostrom (Ed.), *Global Catastrophic Risks* (pp. 308–345). Oxford University Press. Retrieved from https://intelligence.org/files/AIPosNegFactor.pdf

Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics* (No. 2013-1). Machine Intelligence Research Institute. Retrieved from https://intelligence.org/files/IEM.pdf