How We're Predicting AI—or Failing To

Stuart Armstrong
Future of Humanity Institute

Kaj Sotala
Machine Intelligence Research Institute

Abstract

This paper will look at the various predictions that have been made about AI and propose decomposition schemas for analyzing them. It will propose a variety of theoretical tools for analyzing, judging, and improving these predictions. Focusing specifically on timeline predictions (dates given by which we should expect the creation of AI), it will show that there are strong theoretical grounds to expect predictions to be quite poor in this area. Using a database of 95 AI timeline predictions, it will show that these expectations are borne out in practice: expert predictions contradict each other considerably, and are indistinguishable from non-expert predictions and past failed predictions. Predictions that AI lie 15 to 25 years in the future are the most common, from experts and non-experts alike.

NOTE: The findings in this paper are based on a dataset error. For details, see `https://aiimpacts.org/error-in-armstrong-and-sotala-2012/`.

Armstrong, Stuart, and Kaj Sotala. 2012. "How We're Predicting AI—or Failing To."
In *Beyond AI: Artificial Dreams*, edited by Jan Romportl, Pavel Ircing, Eva Zackova, Michal Polak, and Radek Schuster, 52–75. Pilsen: University of West Bohemia.

This version contains minor changes.

1. Introduction

Predictions about the future development of artificial intelligence are as confident as they are diverse. Starting with Turing's initial estimation of a 30% pass rate on Turing test by the year 2000 (Turing 1950), computer scientists, philosophers and journalists have never been shy to offer their own definite prognostics, claiming AI to be impossible (Jacquette 1987) or just around the corner (Darrach 1970) or anything in between.

What are we to make of these predictions? What are they for, and what can we gain from them? Are they to be treated as light entertainment, the equivalent of fact-free editorials about the moral decline of modern living? Or are there some useful truths to be extracted? Can we feel confident that certain categories of experts can be identified, and that their predictions stand out from the rest in terms of reliability?

In this paper, we start off by proposing classification schemes for AI predictions: what types of predictions are being made, and what kinds of arguments or models are being used to justify them. Different models and predictions can result in very different performances, and it will be the ultimate aim of this project to classify and analyze their varying reliability.

Armed with this scheme, we then analyze some of these approaches from the theoretical perspective, seeing whether there are good reasons to believe or disbelieve their results. The aim is not simply to critique individual methods or individuals, but to construct a toolbox of assessment tools that will both enable us to estimate the reliability of a prediction, and allow predictors to come up with better results themselves.

This paper, the first in the project, looks specifically at AI timeline predictions: those predictions that give a date by which we should expect to see an actual AI being developed (we use AI in the old fashioned sense of a machine capable of human-comparable cognitive performance; a less ambiguous modern term would be "AGI," Artificial *General* Intelligence). With the aid of the biases literature, we demonstrate that there are strong reasons to expert that experts would *not* be showing particular skill the field of AI timeline predictions. The task is simply not suited for good expert performance.

Those theoretical results are supplemented with the real meat of the paper: a database of 257 AI predictions, made in a period spanning from the 1950s to the present day. This database was assembled by researchers from the Singularity Institute (Jonathan Wang and Brian Potter) systematically searching though the literature, and is a treasure-trove of interesting results. A total of 95 of these can be considered AI timeline predictions. We assign to each of them a single "median AI" date, which then allows us to demonstrate that AI expert predictions are greatly inconsistent with each other—and indistinguishable from non-expert performance, and past failed predictions.

With the data, we further test two folk theorems: firstly that predictors always predict the arrival of AI just before their own deaths, and secondly that AI is always 15 to 25 years into the future. We find evidence for the second thesis but not for the first.

This enabled us to show that there seems to be no such thing as an "AI expert" for timeline predictions: no category of predictors stands out from the crowd.

2. Taxonomy of Predictions

2.1. Prediction Types

> There will never be a bigger plane built.
>
> —Boeing engineer on the 247, a twin engine plane that held ten people.

The standard image of a prediction is some fortune teller staring deeply into the mists of a crystal ball, and decreeing, with a hideous certainty, the course of the times to come. Or in a more modern version, a scientist predicting the outcome of an experiment or an economist pronouncing on next year's GDP figures. But these "at date X, Y will happen" are just one type of valid prediction. In general, a prediction is something that constrains our expectation of the future. Before hearing the prediction, we thought the future would have certain properties; but after hearing and believing it, we now expect the future to be different from our initial thoughts.

Under this definition, conditional predictions—"if A, then B will happen"—are also perfectly valid. As are negative predictions: we might have believed initially that perpetual motion machines were possible, and imagined what they could be used for. But once we accept that one cannot violate conservation of energy, we have a different picture of the future: one without these wonderful machines and all their fabulous consequences.

For the present analysis, we will divide predictions about AI into four types:

Timelines and outcome predictions. These are the traditional types of predictions, telling us when we will achieve specific AI milestones. Examples: An AI will pass the Turing test by 2000 (Turing 1950); within a decade, AIs will be replacing scientists and other thinking professions (Hall 2011).

Scenarios. These are a type of conditional predictions, claiming that if the conditions of the scenario are met, then certain types of outcomes will follow. Example: If we build a human-level AI that is easy to copy and cheap to run, this will cause mass unemployment among ordinary humans (Hanson 1994).

Plans. These are a specific type of conditional prediction, claiming that if someone decides to implement a specific plan, then they will be successful in achieving a partic-

ular goal. Example: We can build an AI by scanning a human brain and simulating the scan on a computer (Sandberg and Bostrom 2008).

Issues and metastatements. This category covers relevant problems with (some or all) approaches to AI (including sheer impossibility results), and metastatements about the whole field. Examples: an AI cannot be built without a fundamental new understanding of epistemology (Deutsch 2012); generic AIs will have certain (potentially dangerous) behaviors (Omohundro 2008).

There will inevitably be some overlap between the categories, but this division is natural enough for our purposes. In this paper we will be looking at timeline predictions. Thanks to the efforts of Jonathan Wang and Brian Potter at the Singularity Institute, the authors were able to make use of extensive databases of this type of predictions, reaching back from the present day back to the 1950s. Other types of predictions will be analyzed in subsequent papers.

2.2. Prediction Methods

Just as there are many types of predictions, there are many ways of arriving at them—consulting crystal balls, listening to the pronouncements of experts, constructing elaborate models. Our review of published predictions has shown that the prediction methods are far more varied than the types of conclusions arrived at. For the purposes of this analysis, we'll divide the prediction methods into the following loose scheme:

1. Causal models
2. Non-causal models
3. The outside view
4. Philosophical arguments
5. Expert authority
6. Non-expert authority

Causal model are the staple of physics: given certain facts about the situation under consideration (momentum, energy, charge, etc.) a conclusion is reached about what the ultimate state will be. If the facts were different, the end situation would be different.

But causal models are often a luxury outside of the hard sciences, whenever we lack precise understanding of the underlying causes. Some success can be achieved with non-causal models: without understanding what influences what, one can extrapolate trends into the future. Moore's law is a highly successful non-causal model (Moore 1965).

The outside view is a method of predicting that works by gathering together specific examples and claiming that they all follow the same underlying trend. For instance, one

could notice the plethora of Moore's laws across the spectrum of computing (in numbers of transistors, size of hard drives, network capacity, pixels per dollar . . .), note that AI is in the same category, and hence argue that AI development must follow a similarly exponential curve (Kurzweil 1999).

Philosophical arguments are common in the field of AI; some are simple impossibility statements: AI is decreed to be impossible for more or less plausible reasons. But the more thoughtful philosophical arguments point out problems that need to be resolved to achieve AI, highlight interesting approaches to doing so, and point potential issues if this were to be achieved.

Many predictions rely strongly on the status of the predictor: their innate expertise giving them potential insights that cannot be fully captured in their arguments, so we have to trust their judgment. But there are problems in relying on expert opinion, as we shall see.

Finally, some predictions rely on the judgment or opinion of non-experts. Journalists and authors are examples of this, but often actual experts will make claims outside their domain of expertise. CEO's, historians, physicists, and mathematicians will generally be no more accurate than anyone else when talking about AI, no matter how stellar they are in their own field (Kahneman 2011).

Predictions can use a mixture of these approaches, and often do. For instance, Ray Kurzweil's "Law of Time and Chaos" uses the outside view to group together evolutionary development, technological development, and computing into the same category, and constructs a causal model predicting time to the "Singularity" (Kurzweil 1999). Moore's law (non-causal model) is a key input to this Law, and Ray Kurzweil's expertise is the main evidence for the Law's accuracy.

This is the schema we will be using in this paper, and in the prediction databases we have assembled. But the purpose of any such schema is to bring clarity to the analysis, not to force every prediction into a particular box. We hope that the methods and approaches used in this paper will be of general use to everyone wishing to analyze the reliability and usefulness of predictions, in AI and beyond. Hence this schema can be freely adapted or discarded if a particular prediction does not seem to fit it, or if an alternative schema seems to be more useful for the analysis of the question under consideration.

3. A Toolbox of Assessment Methods

The purpose of this paper is not only to assess the accuracy and reliability of some of the AI predictions that have already been made. The purpose is to start building a "toolbox" of assessment methods that can be used more generally, applying them to current and future predictions.

3.1. Extracting Verifiable Predictions

The focus of this paper is squarely on the behavior of AI. This is not a philosophical point; we are not making the logical positivist argument that only empirically verifiable predictions have meaning (Carnap 1928). But it must be noted that many of the vital questions about AI—can it built, when, will it be dangerous, will it replace humans, and so on—all touch upon behavior. This narrow focus has the added advantage that empirically verifiable predictions are (in theory) susceptible to falsification, which means ultimately agreement between people of opposite opinions. Predictions like these have a very different dynamic to those that cannot be shown to be wrong, even in principle.

To that end, we will seek to reduce the prediction to an empirically verifiable format. For some predictions, this is automatic: they are already in the correct format. When Kurzweil wrote "One of my key (and consistent) predictions is that a computer will pass the Turing test by 2029," then there is no need to change anything. Conversely, some philosophical arguments concerning AI, such as some of the variants of the Chinese Room argument (Searle 1980), are argued to contain no verifiable predictions at all: an AI that demonstrated perfect human behavior would not affect the validity of the argument.

And in between there are those predictions that are partially verifiable. Then the verifiable piece must be clearly extracted and articulated. Sometimes it is ambiguity that must be overcome: when an author predicts an AI "Omega point" in 2040 (Schmidhuber 2007), it is necessary to read the paper with care to figure out what counts as an Omega point and (even more importantly) what doesn't.

Even purely philosophical predictions can have (or can be interpreted to have) verifiable predictions. One of the most famous papers on the existence of conscious states is Thomas Nagel's "What is it like to be a bat" (Nagel 1974). In this paper, Nagel argues that bats must have mental states, but that we humans can never understand what it is like to have these mental states. This feels purely philosophical, but does lead to empirical predictions: that if the bat's intelligence were increased and we could develop a common language, then at some point in the conversation with it, our understanding would reach an impasse. We would try to describe what our internal mental states felt like, but would always fail to communicate the essence of our experience to the other species.

Many other philosophical papers can likewise be read as having empirical predictions; as making certain states of the world more likely or less—even if they seem to be devoid of this. The Chinese Room argument, for instance, argues that formal algorithms will lack the consciousness that humans possess (Searle 1980). This may seem to be an entirely self-contained argument—but consider that a lot of human behavior revolves around consciousness, be it discussing it, commenting on it, defining it or in-

tuitively noticing it in others. Hence if we believed the Chinese Room argument, and were confronted with two AI projects, one based on advanced algorithms and one based on modified human brains, we would be likely to believe that the second project is more likely to result in an intelligence that *seemed* conscious than the first. This is simply because we wouldn't believe that the first AI could ever be conscious, and that it is easier to seem conscious when one actually is. And that gives an empirical prediction.

Note that the authors of the predictions may disagree with our "extracted" conclusions. This is not necessarily a game breaker. For instance, even if there is no formal link between the Chinese Room model and the prediction above, it's still the case that the intuitive reasons for believing the model are also good reasons for believing the prediction. Our aim should always be to try to create useful verifiable predictions in any way we can. In this way, we can make use of much more of the AI literature. For instance, Lucas argues that AI is impossible because it could not recognize the truth of its own Gödel sentence (Lucas 1961).[1] This is a very strong conclusion, and we have to accept a lot of Lucas's judgments before we agree with it. Replacing the conclusion with the weaker (and verifiable) "self reference will be an issue with advanced AI, and will have to be dealt with somehow by the programmers" gives us a useful prediction which is more likely to be true.

Care must be taken when applying this method: the point is to extract a useful verifiable prediction, not to weaken or strengthen a reviled or favored argument. The very first stratagems in Shopenhauer's "The Art of Always being Right" (Schopenhauer 1831) are to extend and over-generalize the consequences of your opponent's argument; conversely, one should reduce and narrow down one's own arguments. There is no lack of rhetorical tricks to uphold one's own position, but if one is truly after the truth, one must simply attempt to find the most reasonable empirical version of the argument; the truth-testing will come later.

This method often increases uncertainty, in that it often narrows the consequences of the prediction, and allows more possible futures to exist, consistently with that prediction. For instance, Bruce Edmonds (Edmonds 2008), building on the "No Free Lunch" results (Wolpert and Macready 1995), demonstrates that there is no such thing as a universal intelligence: no intelligence that performs better than average in every circumstance. Initially this seems to rule out AI entirely; but when one analyzes what this means empirically, one realizes there is far less to it. It does not forbid an algorithm from performing better than any human being in any situation any human being would ever

1. A Gödel sentence is a sentence G that can be built in any formal system containing arithmetic. G is implicitly self-referential, as it is equivalent with "there cannot exist a proof of G". By construction, there cannot be a consistent proof of G from within the system.

encounter, for instance. So our initial intuition, which was to rule out all futures with AIs in them, is now replaced by the realization that we have barely put any constraints on the future at all.

3.2. Clarifying and Revealing Assumptions

The previous section was concerned with the predictions' conclusions. Here we will instead be looking at its assumptions, and the logical structure of the argument or model behind it. The objective is to make the prediction as rigorous as possible

Philosophers love doing this: taking apart argument, adding caveats and straightening out the hand-wavy logical leaps. In a certain sense, it can be argued that analytic philosophy is entirely about making arguments rigorous. One of the oldest methods in philosophy—the dialectic (Plato, 380BCE)—also plays this role, with concepts getting clarified during the conversation between philosophers and various Athenians. Though this is perhaps philosophy's greatest contribution to knowledge, it is not exclusively the hunting ground of philosophers. All rational fields of endeavor do—and should!—benefit from this kind of analysis.

Of critical importance is revealing hidden assumptions that went into the predictions. These hidden assumptions—sometimes called Enthymematic gaps in the literature (Fallis 2003)—are very important because they clarify where the true disagreements lie, and where we need to focus our investigation in order to find out the truth of prediction. Too often, competing experts will make broad-based arguments that fly past each other. This makes choosing the right argument a matter of taste, prior opinions and our admiration of the experts involved. But if the argument can be correctly deconstructed, then the source of the disagreement can be isolated, and the issue can be decided on much narrower grounds—and its much clearer whether the various experts have relevant expertise or not (see Section 3.4). The hidden assumptions are often implicit, so it is perfectly permissible to construct assumptions that the predictors were not consciously aware of using.

For example, let's look again at the Gödel arguments mentioned in Section 3.1. The argument shows that formal systems of a certain complexity must be either incomplete (unable to see that their Gödel sentence is true) or inconsistent (proving false statements). This is contrasted with humans, who—allegedly—use meta-reasoning to know that their own Gödel statements are true. It should first be noted here that no one has written down an actual "human Gödel statement," so we cannot be sure humans would actually figure out that it is true.[2] Also, humans are both inconsistent and able to deal

2. One could argue that, by definition, a human Gödel statement must be one that humans cannot recognize as being a human Gödel statement!

with inconsistencies without a complete collapse of logic. In this, they tend to differ from AI systems, though some logic systems such as relevance logic do mimic the same behavior (Routley and Meyer 1976). In contrast, both humans and AIs are not logically omniscient—they are not capable of proving everything provable within their logic system (the fact that there are an infinite number of things to prove being the problem here). So this analysis demonstrates the hidden assumption in Lucas's argument: that the behavior of an actual computer program running on a real machine is more akin to that of a logically omniscient formal agent, than it would be to a real human being. That assumption may be flawed or correct, but is one of the real sources of disagreement over whether Gödelian arguments rule out artificial intelligence.

Again, it needs to be emphasized that the purpose is to clarify and analyze arguments, not to score points for one side or the other. It is easy to phrase assumptions in ways that sound good or bad for either "side." It is also easy to take the exercise too far: finding more and more minor clarifications or specific hidden assumptions until the whole prediction becomes a hundred page mess of over-detailed special cases. The purpose is to clarify the argument until it reaches the point where all (or most) parties could agree that these assumptions are the real sources of disagreement. And then we can consider what empirical evidence, if available, or expert opinion has to say about these disagreements.

There is surprisingly little published on the proper way of clarifying assumptions, making this approach more an art than a science. If the prediction comes from a model, we have some standard tools available for clarifying, though see Morgan and Henrion (1990). Most of these methods work by varying parameters in the model and checking that this doesn't cause a breakdown in the prediction.

3.2.1. Model Testing and Counterfactual Resiliency

Though the above works from inside the model, there are very few methods that can test the strength of a model from the outside. This is especially the case for non-causal models: what are the assumptions behind Moore's famous law (Moore 1965), or Robin Hanson's model that we are due for another technological revolution, based on the timeline of previous revolutions (Hanson 2009)? If we can't extract assumptions, we're reduced to saying "that feel right/wrong to me," and therefore we're getting nowhere.

The authors have come up with a putative way of testing the assumptions of such models (in the case of Moore's law, the empirical evidence in favor is strong, but there is still the question of what is powering the law and whether it will cross over to new chip technologies again and again). It involves giving the model a counterfactual resiliency check: imagining that world history had happened slightly differently, and checking whether the model would have stood up in those circumstances. Counterfactual changes are permitted to anything that the model ignores.

The purpose of this exercise is not to rule out certain models depending on one's own preferred understanding of history (e.g. "Protestantism was essential to the industrial revolution, and was a fluke due to Martin Luther; so it's very likely that the industrial revolution would not have happened in the way or timeframe that it did, hence Hanson's model—which posits the industrial revolutions's dates as inevitable—is wrong"). Instead it is to illustrate the tension between the given model and other models of history (e.g. "The assumptions that Protestantism was both a fluke and essential to the industrial revolution are in contradiction with Hanson's model. Hence Hanson's model implies that either Protestantism was inevitable or that it was non-essential to the industrial revolution, a extra hidden assumption"). The counterfactual resiliency exercise has been carried out at length in an online post (Armstrong 2012). The general verdict seemed to be that Hanson's model contradicted a lot of seemingly plausible assumptions about technological and social development. Moore's law, on the other hand, seemed mainly dependent on the continuing existence of a market economy and the absence of major catastrophes.

This method is new, and will certainly be refined in future. Again, the purpose of the method is not to rule out certain models, but to find the nodes of disagreement.

3.2.2. More Uncertainty

Clarifying assumptions often ends up increasing uncertainty, as does revealing hidden assumptions. The previous section focused on extracting verifiable predictions, which often increases the range of possible worlds compatible with a prediction. Here, by clarifying and caveating assumptions, and revealing hidden assumption, we reduce the number of worlds in which the prediction is valid. This means that the prediction puts fewer constraints on our expectations. In counterpart, of course, the caveated prediction is more likely to be true.

3.3. Empirical Evidence

The gold standard in separating true predictions from false ones must always be empirical evidence. The scientific method has proved to be the best way of disproving false hypotheses, and should be used whenever possible. Other methods, such as expert opinion or unjustified models, come nowhere close.

The problem with empirical evidence is that . . . it is generally non-existent in the AI prediction field. Since AI predictions are all about the existence and properties of a machine that hasn't yet been built, that no-one knows how to build or whether it actually can be built, there is little opportunity for the whole hypothesis-prediction-testing cycle. This should indicate the great difficulties in the field. Social sciences, for instance, are often seen as the weaker cousins of the hard sciences, with predictions much

more contentious and less reliable. And yet the social sciences make use of the scientific method, and have access to some types of repeatable experiments. Thus any prediction in the field of AI should be treated as less likely than any social science prediction.

That generalization is somewhat over-harsh. Some AI prediction methods hew closer to the scientific method, such as the whole brain emulations model (Sandberg and Bostrom 2008)—it makes testable predictions along the way. Moore's law is a wildly successful prediction, and connected to some extent with AI. Many predictors (e.g. Kurzweil) make partial predictions on the road towards AI; these can and should be assessed—track records allow us to give some evidence to the proposition "this expert knows what they're talking about." And some models also allow for a degree of testing. So the field is not void of empirical evidence; it's just that there is so little of it, and to a large extent we must put our trust in expert opinion.

3.4. Expert Opinion

Reliance on experts is nearly unavoidable in AI prediction. Timeline predictions are often explicitly based on experts' feelings; even those that consider factors about the world (such as computer speed) need an expert judgment about why that factor is considered and not others. Plans need experts to come up with them and judge their credibility. And unless every philosopher agrees on the correctness of a particular philosophical argument, we are dependent to some degree on the philosophical judgment of the author. It is the purpose of all the methods described above that we can refine and caveat a prediction, back it up with empirical evidence whenever possible, and thus clearly highlight the points where we need to rely on expert opinion. And so can focus on the last remaining points of disagreement: the premises themselves (that is of course the ideal situation: some predictions are given directly with no other basis but expert authority, meaning there is nothing to refine).

Should we expect experts to be good at this task? There have been several projects over the last few decades to establish the domains and tasks where we would expect experts to have good performance (Shanteau 1992; Kahneman and Klein 2009). Table 1 summarizes the results:

Not all of these are directly applicable to the current paper (are predictions about human level AIs predictions about things, or about behavior?). One of the most important factors is whether experts get feedback, preferably immediate feedback. We should expect the best expert performance when their guesses are immediately confirmed or disconfirmed. When feedback is unavailable or delayed, or the environment isn't one that give good feedback, then expert performance drops precipitously (Kahneman and Klein 2009; Kahneman 2011).

Table 1: Table of task properties conducive to good and poor expert performance.

Good performance:	Poor performance:
Static stimuli	Dynamic (changeable) stimuli
Decisions about things	Decisions about behavior
Experts agree on stimuli	Experts disagree on stimuli
More predictable problems	Less predictable problems
Some errors expected	Few errors expected
Repetitive tasks	Unique tasks
Feedback available	Feedback unavailable
Objective analysis available	Subjective analysis only
Problem decomposable	Problem not decomposable
Decision aids common	Decision aids rare

Table 1 applies to both domain and task. Any domain of expertise strongly in the right column will be one where we expect poor expert performance. But if the individual expert tries to move their own predictions into the left column (maybe by decomposing the problem as far as it will go, training themselves on related tasks where feedback is available . . .) they will be expected to perform better. In general, we should encourage this type of approach.

When experts fail, there are often simple algorithmic models that demonstrate better performance (Grove et al. 2000). In these cases, the experts often just spell out their criteria, design the model in consequence, and let the model give its predictions: this results in better predictions than simply asking the expert in the first place. Hence we should also be on the lookout for experts who present their findings in the form of a model.

As everyone knows, experts sometimes disagree. This fact strikes at the very heart of their supposed expertise. We listen to them because they have the skills and experience to develop correct insights. If other experts have gone through the same process and come to an opposite conclusion, then we have to conclude that their insights do not derive from their skills and experience, and hence should be discounted. Now if one expert opinion is a fringe position held by only a few experts, we may be justified in dismissing it simply as an error. But if there are different positions held by large numbers of disagreeing experts, how are we to decide between them? We need some sort of objective criteria: we are not experts in choosing between experts, so we have no special skills in deciding the truths on these sorts of controversial positions.

What kind of objective criteria could there be? A good track record can be an indicator, as is a willingness to make verifiable, non-ambiguous predictions. A better connection with empirical knowledge and less theoretical rigidity are also positive indications (Tetlock 2005), and any expert that approached their task with methods that were more on the left of the table than on the right should be expected to be more correct. But these are second order phenomena—we're looking at our subjective interpretation of expert's subjective opinion—so in most cases, when there are strong disagreement between experts, we simply can't tell which position is true.

3.4.1. Grind Versus Insight

Some AI prediction claim that AI will result from grind: i.e. lots of hard work and money. Other claim that AI will need special insights: new unexpected ideas that will blow the field wide open (Deutsch 2012).

In general, we are quite good at predicting grind. Project managers and various leaders are often quite good at estimating the length of projects (as long as they're not directly involved in the project (Buehler, Griffin, and Ross 1994)). Even for relatively creative work, people have sufficient feedback to hazard reasonable guesses. Publication dates for video games, for instance, though often over-optimistic, are generally not ridiculously erroneous—even though video games involve a lot of creative design, play-testing, art, programing the game "AI," etc. Moore's law could be taken as an ultimate example of grind: we expect the global efforts of many engineers across many fields to average out to a rather predictable exponential growth.

Predicting insight, on the other hand, seems a much more daunting task. Take the Riemann hypothesis, a well-established mathematical hypothesis from 1885, (Riemann 1859). How would one go about estimating how long it would take to solve? How about the $P = NP$ hypothesis in computing? Mathematicians seldom try and predict when major problems will be solved, because they recognize that insight is very hard to predict. And even if predictions could be attempted (the age of the Riemann's hypothesis hints that it probably isn't right on the cusp of being solved), they would need much larger error bars than grind predictions. If AI requires insights, we are also handicapped by the fact of not knowing what these insights are (unlike the Riemann hypothesis, where the hypothesis is clearly stated, and only the proof is missing). This could be mitigated somewhat if we assumed there were several different insights, each of which could separately lead to AI. But we would need good grounds to assume that.

Does this mean that in general predictions that are modeling grind should be accepted more than predictions that are modeling insight? Not at all. Predictions that are modeling grind should only be accepted if they can make a good case that producing an AI is a matter grind only. The predictions around whole brain emulations (Sandberg

and Bostrom 2008), are one of the few that make this case convincingly; this will be analyzed in a subsequent paper.

3.4.2. Non-expert Opinion

It should be borne in mind that all the caveats and problems with expert opinion apply just as well to non-experts. With one crucial difference: we have no reason to trust the non-expert's opinion in the first place. That is not to say that non-experts cannot come up with good models, convincing timelines, or interesting plans and scenarios. It just means that our assessment of the quality of the prediction depends only on what we are given; we cannot extend a non-expert any leeway to cover up a weak premise or a faulty logical step. To ensure this, we should try and assess non-expert predictions blind, without knowing who the author is. If we can't blind them, we can try and get a similar effect by asking ourselves hypothetical questions such as: "Would I find this prediction more or less convincing if the author was the Archbishop of Canterbury? What if it was Warren Buffet? Or the Unabomber?" We should aim to reach the point where hypothetical changes in authorship do not affect our estimation of the prediction.

4. Timeline Predictions

The practical focus of this paper is on AI timeline predictions: predictions giving dates for AIs with human-comparable cognitive abilities. Researchers from the Singularity Institute have assembled a database of 257 AI predictions since 1950, of which 95 include AI timelines.

4.1. Subjective Assessment

A brief glance at Table 1 allows us to expect that AI timeline predictions will generally be of very poor quality. The only factor that is unambiguously positive for AI predictions is that prediction errors are expected and allowed: apart from that, the task seems singularly difficult, especially on the key issue of feedback. An artificial intelligence is a hypothetical machine, which has never existed on this planet before and about whose properties we have but the haziest impression. Most AI experts will receive no feedback whatsoever about their predictions, meaning they have to construct them entirely based on their untested impressions.

There is nothing stopping experts from decomposing the problem, or constructing models which they then calibrate with available data, or putting up interim predictions to test their assessment. And some do use these better approaches (see for instance (Kurzweil 1999; Hanson 1994; Waltz 1988)). But a surprisingly large number don't!

Some predictions are unabashedly based simply on the feelings of the predictor (Good 1962; Armstrong 2007).

Yet another category are of the "Moore's law hence AI" type. They postulate that AI will happen when computers reach some key level, often comparing with some key property of the brain (number of operations per second (Bostrom 1998), or neurones/synapses[3]). In the division established in Section 3.4.1, this is pure "grind" argument: AI will happen after a certain amount of work is performed. But, as we saw, these kinds of arguments are only valid if the predictor has shown that reaching AI does not require new insights! And that step is often absent from the argument.

4.2. Timeline Prediction Data

The above were subjective impressions, formed while looking over the whole database. To enable more rigorous analysis, the various timeline predictions were reduced to a single number for purposes of comparison: this would be the date upon which the predictor expected "human level AI" to be developed.

Unfortunately not all the predictions were in the same format. Some gave ranges, some gave median estimates, some talked about superintelligent AI, others about slightly below-human AI. In order to make the numbers comparable, one of the authors (Stuart Armstrong) went through the list and reduced the various estimates to a single number. He followed the following procedure to extract a "Median human-level AI estimate":

When a range was given, he took the mid-point of that range (rounded down). If a year was given with a 50% likelihood estimate, he took that year. If it was the collection of a variety of expert opinions, he took the prediction of the median expert. If the predictor foresaw some sort of AI by a given date (partial AI or superintelligent AI), and gave no other estimate, he took that date as their estimate rather than trying to correct it in one direction or the other (there were roughly the same number of subhuman AIs as suphuman AIs in the list, and not that many of either). He read extracts of the papers to make judgement calls when interpreting problematic statements like "within thirty years" or "during this century" (is that a range or an end-date?). Every date selected was either an actual date given by the predictor, or the midpoint of a range.[4]

It was also useful to distinguish between popular estimates, performed by journalists, writers or amateurs, from those predictions done by those with expertise in relevant fields (AI research, computer software development, etc.). Thus each prediction was

3. See for instance Dani Eder's 1994 Newgroup posting `http://www.aleph.se/Trans/Global/Singularity/singul.txt`

4. The data can be found at `http://www.neweuropeancentury.org/SIAI-FHI_AI_predictions.xls`; readers are encouraged to come up with their own median estimates.

noted as "expert" or "non-expert"; the expectation being that experts would demonstrate improved performance over non-experts.

Figure 1 graphs the results of this exercise (the range has been reduced; there were seven predictions setting dates beyond the year 2100, three of them expert.)

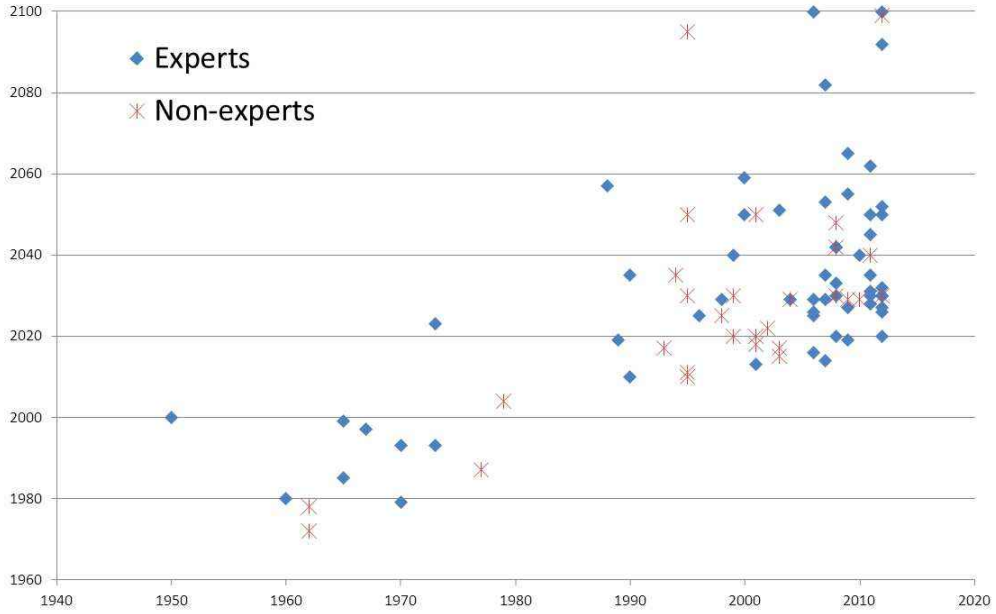

Figure 1: Median estimate for human-level AI, graphed against date of prediction.

As can be seen, expert predictions span the whole range of possibilities and seem to have little correlation with each other. The range is so wide—fifty year gaps between predictions are common—that it provides strong evidence that experts are not providing good predictions. There does not seem to be any visible difference between expert and non-expert performance either, suggesting that the same types of reasoning may be used in both situations, thus negating the point of expertise.

Two explanations have been generally advanced to explain poor expert performance in these matters. The first, the so-called Maes-Garreau law[5] posits that AI experts predict AI happening towards the end of their own lifetime. This would make AI into a technology that would save them from their own deaths, akin to a "Rapture of the Nerds."

The second explanation is that AI is perpetually fifteen to twenty-five years into the future. In this way (so the explanation goes), the predictor can gain credit for working on something that will be of relevance, but without any possibility that their prediction could be shown to be false within their current career.

5. Kevin Kelly, editor of Wired magazine, created the law in 2007 after being influenced by Pattie Maes at MIT and Joel Garreau (author of Radical Evolution).

We'll now look at the evidence for these two explanations.

4.2.1. Nerds Don't Get Raptured

Fifty-five predictions were retained, in which it was possible to estimate the predictor's expected lifespan. Then the difference between their median prediction and this lifespan was computed (a positive difference meaning they would expect to die before AI, a negative difference meaning they didn't). A zero difference would be a perfect example of the Maes-Garreau law: the predictor expects AI to be developed at the exact end of their life. This number was then plotted again the predictor's age in Figure 2 (the plot was restricted to those predictions within thirty years of the predictor's expected lifetime).

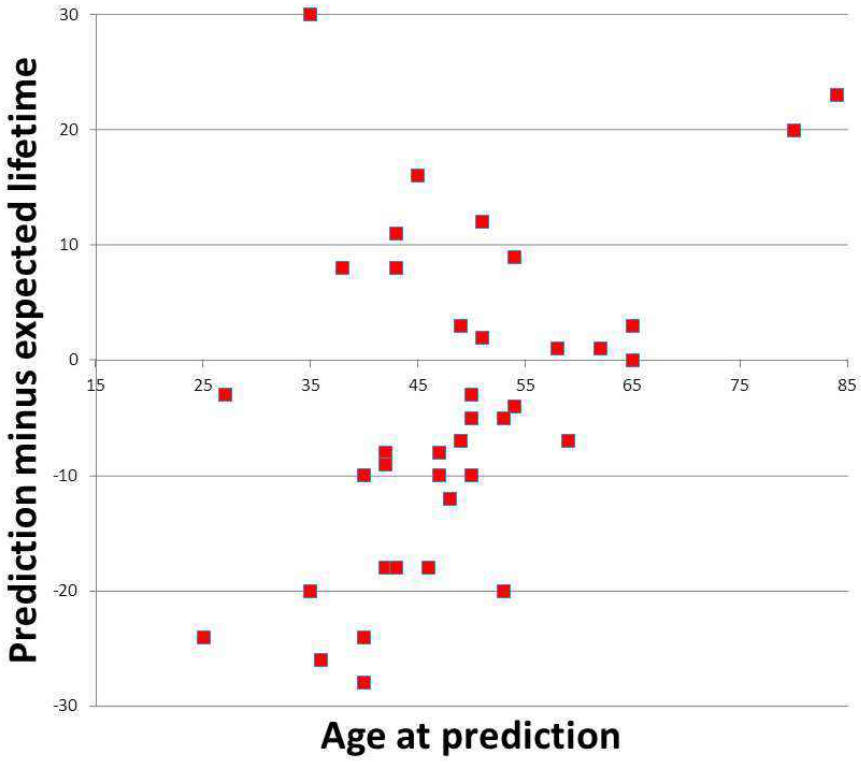

Figure 2: Difference between the predicted time to AI and the predictor's life expectancy, graphed against the predictor's age.

From this, it can be seen that the Maes-Garreau law is not borne out by the evidence: only twelve predictions (22% of the total) were within five years in either direction of the zero point.

4.2.2. Twenty Years to AI

The "time to AI" was computed for each expert prediction. This was graphed in Figure 3. This demonstrates a definite increase in the 16–25 year predictions: 21 of the 62 expert

predictions were in that range (34%). This can be considered weak evidence that experts do indeed prefer to predict AI happening in that range from their own time.

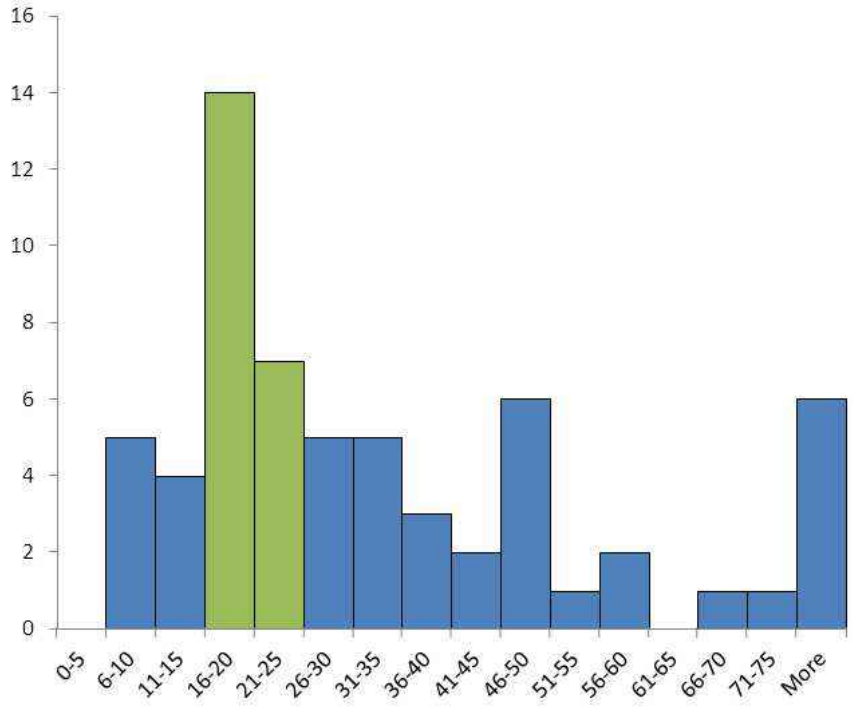

Figure 3: Time between the arrival of AI and the date the prediction was made, for expert predictors.

But the picture gets more damning when we do the same plot for the non-experts, as in Figure 4. Here, 13 of the 33 predictions are in the 16-25 year range. But more disturbingly, the time to AI graph is almost identical for experts and non-experts! Though this does not preclude the possibility of experts being more accurate, it does hint strongly that experts and non-experts may be using similar psychological procedures when creating their estimates.

The next step is to look at failed predictions. There are 15 of those, most dating to before the "AI winter" in the eighties and nineties. These have been graphed in Figure 5—and there is an uncanny similarity with the other two graphs! So expert predictions are not only indistinguishable from non-expert predictions, they are also indistinguishable from past failed predictions. Hence it is not unlikely that recent predictions are suffering from the same biases and errors as their predecessors

5. Conclusion

This paper, the first in a series analyzing AI predictions, focused on the reliability of AI timeline predictions (predicting the dates upon which "human-level" AI would be de-

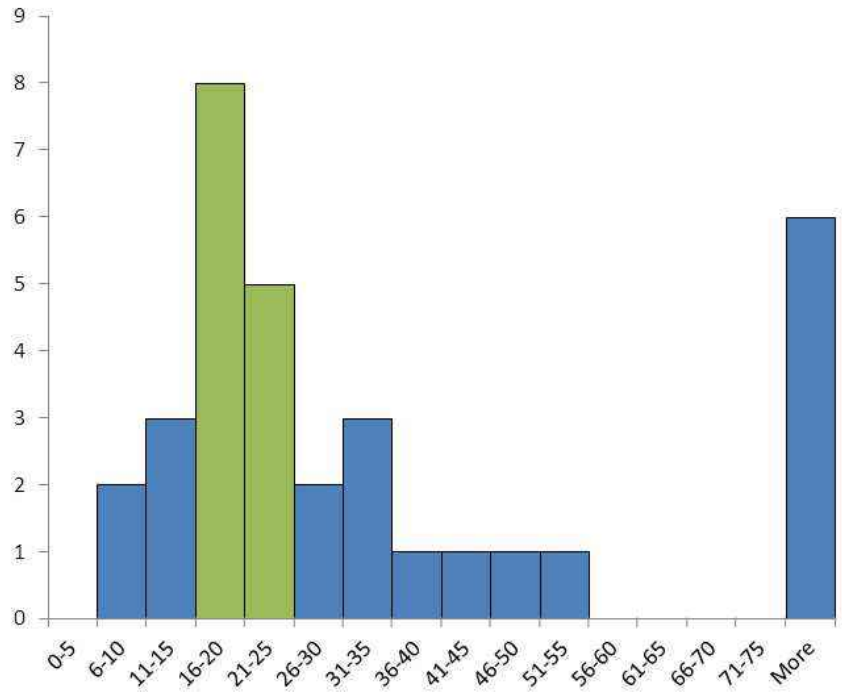

Figure 4: Time between the arrival of AI and the date the prediction was made, for non-expert predictors.

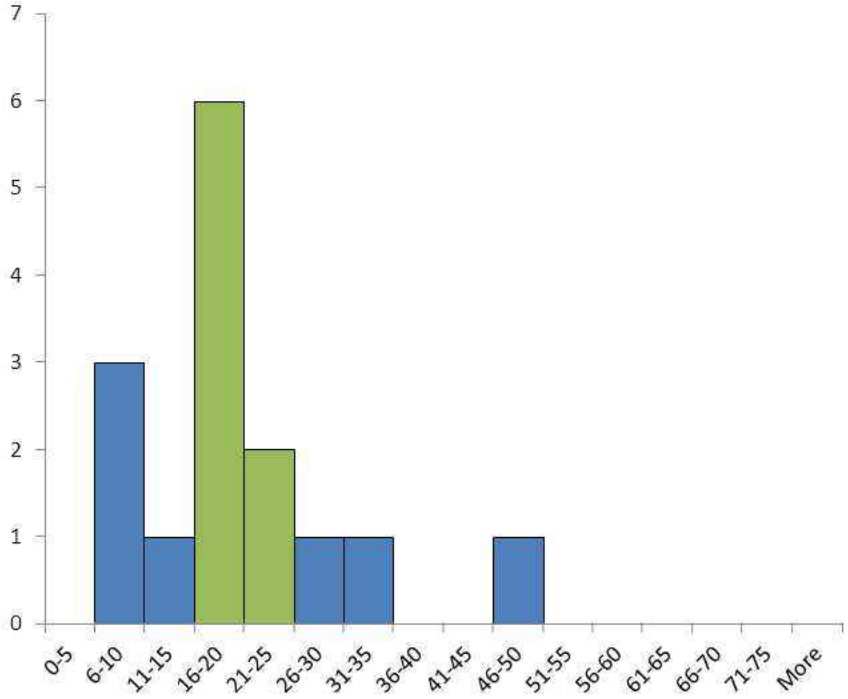

Figure 5: Time between the arrival of AI and the date the prediction was made, for failed predictions.

veloped). These predictions are almost wholly grounded on expert judgment. The biases literature classified the types of tasks on which experts would have good performance, and AI timeline predictions have all the hallmarks of tasks on which they would perform badly.

This was borne out by the analysis of 95 timeline predictions in the database assembled by the Singularity Institute. There were strong indications therein that experts performed badly. Not only were expert predictions spread across a wide range and in strong disagreement with each other, but there was evidence that experts were systematically preferring a "15 to 25 years into the future" prediction. In this, they were indistinguishable from non-experts, and from past predictions that are known to have failed. There is thus no indication that experts brought any added value when it comes to estimating AI timelines. On the other hand, another theory—that experts were systematically predicting AI arrival just before the end of their own lifetime—was seen to be false in the data we have.

There is thus strong grounds for dramatically increasing the uncertainty in any AI timeline prediction.

Acknowledgments

The authors wish to acknowledge the help and support of the Singularity Institute, the Future of Humanity Institute and the James Martin School, as well as the individual advice of Nick Bostrom, Luke Muelhauser, Vincent Mueller, Anders Sandberg, Lisa Makros, Sean O'Heigeartaigh, Daniel Dewey, Eric Drexler and the online community of Less Wrong.

References

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Accessed December 31, 2012. http://www.neweuropeancentury.org/GodAI.pdf.

———. 2012. "Counterfactual Resiliency Test for Non-Causal Models." *Less Wrong* (blog), August 30. http://lesswrong.com/lw/ea8/counterfactual_resiliency_test_for_noncausal/.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

Buehler, Roger, Dale Griffin, and Michael Ross. 1994. "Exploring the 'Planning Fallacy': Why People Underestimate Their Task Completion Times." *Journal of Personality and Social Psychology* 67 (3): 366–381.

Carnap, Rudolf. 1928. *Der Logische Aufbau der Welt* [The logical structure of the world]. Berlin-Schlachtensee: Weltkreis.

Darrach, Brad. 1970. "Meet Shakey, the First Electronic Person." *Life,* November 20, 58–68.

Deutsch, David. 2012. "The Very Laws of Physics Imply That Artificial Intelligence Must Be Possible. What's Holding Us Up?" *Aeon,* October 3. Accessed December 6, 2012. http://www.aeonmagazine.com/being-human/david-deutsch-artificial-intelligence/.

Edmonds, Bruce. 2008. "The Social Embedding of Intelligence: Towards Producing a Machine That Could Pass the Turing Test." In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer,* edited by Robert Epstein, Gary Roberts, and Grace Beber, 211–235. New York: Springer.

Fallis, Don. 2003. "Intentional Gaps In Mathematical Proofs." *Synthese* 134 (1–2): 45–69.

Good, Irving John. 1962. *The Scientist Speculates: An Anthology of Partly-Baked Ideas.* Edited by Alan James Mayne and John Maynard Smith. New York: Basic Books.

Grove, William M., David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. 2000. "Clinical Versus Mechanical Prediction: A Meta-Analysis." *Psychological Assessment* 12 (1): 19–30.

Hall, John Storrs. 2011. *Further Reflections on the Timescale of AI.* Melbourn, Australia, November 30. Paper presented at the Solomonoff 85th Memorial Conference. http://www.solomonoff85thmemorial.monash.edu/accepted_papers.html.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

———. 2009. "The Economics of Brain Emulations." In *Unnatural Selection: The Challenges of Engineering Tomorrow's People,* edited by Peter Healey and Steve Rayner. Science in Society. Sterling, VA: Earthscan.

Jacquette, Dale. 1987. "Metamathematical Criteria for Minds and Machines." *Erkenntnis* 27 (1): 1–16.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* 1st ed. New York: Farrar, Straus / Giroux.

Kahneman, Daniel, and Gary Klein. 2009. "Conditions for Intuitive Expertise: A Failure to Disagree." *American Psychologist* 64 (6): 515–526.

Kurzweil, Ray. 1999. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence.* New York: Viking.

Lucas, J. R. 1961. "Minds, Machines and Gödel." *Philosophy* 36 (137): 112–127.

Moore, Gordon. 1965. "Cramming More Components onto Integrated Circuits." *Electronics* 38 (8): 114–117. `http://download.intel.com/museum/Moores_Law/Articles-Press_Releases/Gordon_Moore_1965_Article.pdf`.

Morgan, M. Granger, and Max Henrion. 1990. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* New York: Cambridge University Press.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (4): 435–450. `http://www.jstor.org/stable/2183914`.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Plato. 380BCE. *The Republic.* Translated by Benjamin Jowett. Republication of 1894 standard text. Project Gutenberg. `http://www.gutenberg.org/files/1497/1497-h/1497-h.htm`.

Riemann, Bernhard. 1859. "Ueber die Anzahl der Primzahlen unter einer gegebenen Grösse" [On the number of primes less than a given magnitude]. *Monatsberichte der Berliner Akademie* (November): 671–680.

Routley, Richard, and Robert K. Meyer. 1976. "Dialectical Logic, Classical Logic, and the Consistency of the World." *Studies in Soviet Thought* 16 (1–2): 1–25.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. `http://www.fhi.ox.ac.uk/wp-content/uploads/brain-emulation-roadmap-report1.pdf`.

Schmidhuber, Jürgen. 2007. "The New AI: General & Sound & Relevant for Physics." In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio Pennachin, 175–198. Cognitive Technologies. Berlin: Springer.

Schopenhauer, Arthur. 1831. *The Art of Being Right: 38 Ways to Win an Argument* [Eristische dialektik: die kunst, recht zu behalten]. Translated by Thomas Bailey Saunders. Republication of 1896 translation. Wikisource. `http://en.wikisource.org/wiki/The_Art_of_Being_Right`.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (03): 417–424.

Shanteau, James. 1992. "Competence in Experts: The Role of Task Characteristics." *Organizational Behavior and Human Decision Processes* 53 (2): 252–266.

Tetlock, Philip E. 2005. *Expert Political Judgment: How Good is it? How Can We Know?* Princeton, NJ: Princeton University Press.

Turing, A. M. 1950. "Computing Machinery and Intelligence." *Mind* 59 (236): 433–460.

Waltz, David L. 1988. "The Prospects for Building Truly Intelligent Machines." In "Artificial Intelligence," *Daedalus* 117 (1): 191–212. http://www.jstor.org/stable/20025144.

Wolpert, David H., and William G. Macready. 1995. *No Free Lunch Theorems for Search.* Santa Fe, NM: The Santa Fe Institite, February 6. http://www.santafe.edu/media/workingpapers/95-02-010.pdf.