Mind and Machines

Session 2: The Future of Intelligence

Optional Pre-Reading

Index

Part I: What does intelligence mean, and how has it changed?

1.	**Computing Machinery and Intelligence** A. M. Turing, Mind 59(236): 433–460 · October 1950	3
2.	**A Formal Measure of Machine Intelligence** Shane Legg and Marcus Hutter, IDSIA Technical Report IDSIA-10-06 · 14 April 2006	5
3.	**Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data** Emily M. Bender and Alexander Koller, Proceedings of the 58th Annual Meeting of the ACL, pp. 5185–5198 · July 2020	8

Part II: What would a superintelligent AI look like?

4.	**The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents** Nick Bostrom, Minds and Machines 22(2): 71–85 · 2012	17
5.	**The AI Cargo Cult: The Myth of a Superhuman AI** Kevin Kelly, Backchannel · 25 April 2017	26
6.	**Machines of Loving Grace: How AI Could Transform the World for the Better** Dario Amodei · October 2024	32

Other Links

7.	**What happens when our computers get smarter than we are?** Nick Bostrom, TED2015 (talk and transcript) · March 2015	34
8.	**How Long Before Superintelligence?** Nick Bostrom, International Journal of Futures Studies 2 · 1998	34
9.	**Superintelligence: Paths, Dangers, Strategies** Nick Bostrom, Oxford University Press · 2014	34
10.	**The implausibility of intelligence explosion** François Chollet, Medium · 27 November 2017	34
11.	**On the Measure of Intelligence** François Chollet, arXiv:1911.01547 · November 2019	34
12.	**A Collection of Definitions of Intelligence** Shane Legg and Marcus Hutter, arXiv:0706.3639 · June 2007	34
13.	**The Singularity: A Philosophical Analysis** David J. Chalmers, Journal of Consciousness Studies 17: 7–65 · 2010	34
14.	**Minds, Brains, and Programs** John R. Searle, Behavioral and Brain Sciences 3(3): 417–457 · 1980	34
15.	**Why AI is Harder Than We Think** Melanie Mitchell, arXiv:2104.12871 · April 2021	34
16.	**The Gentle Singularity** Sam Altman · 10 June 2025	34
17.	**AI 2027** Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland and Romeo Dean · 3 April 2025	34

Computing Machinery and Intelligence

A. M. Turing, Mind 59(236): 433–460 · October 1950
Extract: §1 The Imitation Game, pp. 433–434
https://doi.org/10.1093/mind/LIX.236.433

VOL. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

Now suppose X is actually A, then A must answer. It is A's

object in the game to try and cause C to make the wrong identification. His answer might therefore be

'My hair is shingled, and the longest strands are about nine inches long.'

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as "I am the woman, don't listen to him!' to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'

2. *Critique of the New Problem.*

As well as asking, 'What is the answer to this new form of the question', one may ask, 'Is this new question a worthy one to investigate?' This latter question we investigate without further ado, thereby cutting short an infinite regress.

The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a 'thinking machine' more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices. Some other advantages of the proposed criterion may be shown up by specimen questions and answers. Thus:

Q: Please write me a sonnet on the subject of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764
A: (Pause about 30 seconds and then give as answer) 105621.
Q: Do you play chess?
A: Yes.

A Formal Measure of Machine Intelligence

Shane Legg and Marcus Hutter, IDSIA Technical Report IDSIA-10-06 · 14 April 2006

Extract: §1–3, pp. 1–3

https://arxiv.org/abs/cs/0605024

arXiv:cs/0605024v1 [cs.AI] 6 May 2006

Technical Report IDSIA-10-06

A Formal Measure of Machine Intelligence

Shane Legg and **Marcus Hutter**

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
{shane,marcus}@idsia.ch http://www.idsia.ch/

14 April 2006

Abstract

A fundamental problem in artificial intelligence is that nobody really knows what intelligence is. The problem is especially acute when we need to consider artificial systems which are significantly different to humans. In this paper we approach this problem in the following way: We take a number of well known informal definitions of human intelligence that have been given by experts, and extract their essential features. These are then mathematically formalised to produce a general measure of intelligence for arbitrary machines. We believe that this measure formally captures the concept of machine intelligence in the broadest reasonable sense.

1 Introduction

Most of us think that we recognise intelligence when we see it, but we are not really sure how to precisely define or measure it. We informally judge the intelligence of others by relying on our past experiences in dealing with people. Naturally, this naive approach is highly subjective and imprecise. A more principled approach would be to use one of the many standard intelligence tests that are available. Contrary to popular wisdom, these tests, when correctly applied by a professional, deliver statistically consistent results and have considerable power to predict the future performance of individuals in many mentally demanding tasks. However, while these tests work well for humans, if we wish to measure the intelligence of other things, perhaps of a monkey or a new machine learning algorithm, they are clearly inappropriate.

One response to this problem might be to develop specific kinds of tests for specific kinds of entities; just as intelligence tests for children differ to intelligence tests for adults. While this works well when testing humans of different ages, it comes undone when we need to measure the intelligence of entities which are profoundly different to each other in terms of their cognitive capacities, speed, senses, environments in which they operate, and so on. To measure the intelligence of such diverse systems in a meaningful way we must step back from the specifics of particular systems and establish the underlying fundamentals of what it is that we are really trying to measure. That is, we need to establish a notion of intelligence that goes beyond the specifics of particular kinds of systems.

The difficulty of doing this is readily apparent. Consider, for example, the memory and numerical computation tasks that appear in some intelligence tests and which were once regarded as defining hallmarks of human intelligence. We now know that these tasks are absolutely trivial for a machine and thus do not test the machine's intelligence. Indeed even the mentally demanding task of playing chess has been largely reduced to brute force search. As technology advances, our concept of what intelligence is continues to evolve with it.

How then are we to develop a concept of intelligence that is applicable to all kinds of systems? Any proposed definition must encompass the essence of human intelligence, as well as other possibilities, in a consistent way. It should not be limited to any particular set of senses, environments or goals, nor should it be limited to any specific kind of hardware, such as silicon or biological neurons. It should be based on principles which are sufficiently fundamental so as to be unlikely to alter over time. Furthermore, the intelligence measure should ideally be formally expressed, objective, and practically realisable.

This paper approaches this problem in the following way. In *Section 2* we consider a range of definitions of human intelligence that have been put forward by well known psychologists. From these we extract the most common and essential features and use them to create an informal definition of intelligence. *Section 3* then introduces the frame-

work which we use to construct our formal measure of intelligence. This framework is formally defined in *Section 4*. In *Section 5* we use our developed formalism to produce a formal definition of intelligence. *Section 7* closes with a short summary.

A preliminary sketch of the ideas in this paper appeared in the poster [LH05]. It can be shown that the intelligence measure presented here is in fact a variant of the Intelligence Order Relation that appears in the theory of AIXI, the provably optimal universal agent [Hut04]. A long journal version of this paper is being written in which we give the proposed measure of machine intelligence and its relation to other such tests a much more comprehensive treatment.

Naturally, we expect such a bold initiative to be met with resistance. However, we hope that the reader will appreciate the value of our approach: With a formally precise definition put forward we aim to better our understanding of what is a notoriously subjective and slippery concept.

2 The concept of intelligence

Although definitions of human intelligence given by experts in the field vary, most of their views cluster around a few common perspectives. Perhaps the most common perspective, roughly stated, is to think of intelligence as being the ability to successfully operate in uncertain environments by learning and adapting based on experience. The following often quoted definitions, which can be found in [Ste00], [Wec58], [Bin37] and [Got97], all express this notion of intelligence but with different emphasis in each case:

- "The capacity to learn or to profit by experience." – W. F. Dearborn
- "Ability to adapt oneself adequately to relatively new situations in life." – R. Pinter
- "A person possesses intelligence insofar as he has learned, or can learn, to adjust himself to his environment." – S. S. Colvin
- "We shall use the term 'intelligence' to mean the ability of an organism to solve new problems. . . ." – W. V. Bingham
- "A global concept that involves an individual's ability to act purposefully, think rationally, and deal effectively with the environment." – D. Wechsler
- "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience." – L. S. Gottfredson and 52 expert signatories

These definitions have certain common features; in some cases they are explicitly stated, while in others they are more implicit. Perhaps the most elementary feature is that intelligence is seen as a property of an entity which is interacting with an external environment, problem or situation. Indeed this much is common to practically all proposed definitions of intelligence. As we will be referring back to these concepts regularly, we will refer to the entity whose intelligence is in question as the *agent*, and the external environment, problem or situation that it faces as the *environment*. An environment could be a large complex world in which the agent exists, similar to the usual meaning, or something as narrow as a game of tic-tac-toe.

The second common feature of these definitions is that an agent's intelligence is related to its ability to succeed in an environment. This implies that the agent has some kind of an objective. Perhaps we could consider an agent intelligent, in an abstract sense, without having any objective. However without any objective what so ever, the agent's intelligence would have no observable consequences. Intelligence then, at least the concrete kind that interests us, comes into effect when an agent has an objective to apply its intelligence to. Here we will refer to this as its *goal*.

The emphasis on learning, adaption and experience in these definitions implies that the environment is not fully known to the agent and may contain surprises and new situations which could not have been anticipated in advance. Thus intelligence is not the ability to deal with one fixed and known environment, but rather the ability to deal with some range of possibilities which cannot be wholly anticipated. This means that an intelligent agent may not be the best possible in any specific environment, particularly before it has had sufficient time to learn. What is important is that the agent is able to learn and adapt so as to perform well over a wide range of specific environments.

Although there is a great deal more to this topic than we have presented here, the above brief analysis gives us the necessary building blocks for our informal working definition of intelligence:

> *Intelligence measures an agent's ability to achieve goals in a wide range of environments.*

We realise that some researchers who study intelligence will take issue with this definition. Given the diversity of views on the nature of intelligence,

a debate which is still being fought, this is unavoidable. Nevertheless, we are confident that our proposed informal working definition is fairly mainstream. We also believe that our definition captures what we are interested in achieving in machines: A very general and flexible capacity to succeed when faced with a wide range of problems and situations. Even those who subscribe to different perspectives on the nature and correct definition of intelligence will surely agree that this is a central objective for anyone wishing to extend the power and usefulness of machines. It is also a definition that can be successfully formalised.

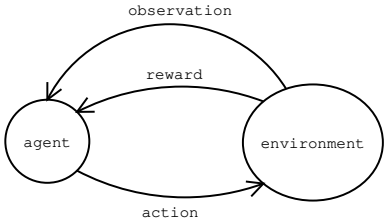


Figure 1: The agent and the environment interact by sending action, observation and reward signals to each other.

3 The agent-environment framework

In the previous section we identified three essential components for our model of intelligence: An agent, an environment, and a goal. Clearly, the agent and the environment must be able to interact with each other; specifically, the agent needs to be able to send signals to the environment and also receive signals being sent from the environment. Similarly the environment must be able to receive and send signals to the agent. In our terminology we will adopt the agent's perspective on these communications and refer to the signals from the agent to the environment as *actions*, and the signals from the environment as *perceptions*.

What is missing from this setup is the goal. As discussed in the previous section, our definition of an agent's intelligence requires there to be some kind of goal for the agent to try to achieve. This implies that the agent somehow knows what the goal is. One possibility would be for the goal to be known in advance and for this knowledge to be built into the agent. The problem with this however is that it limits each agent to just one goal. We need to allow agents which are more flexible than this.

If the goal is not known in advance, the other alternative is to somehow inform the agent of what the goal is. For humans this is easily done using language. In general however, the possession of a sufficiently high level of language is too strong an assumption to make about the agent. Indeed, even for something as intelligent as a dog or a cat, direct explanation will obviously not work.

Fortunately there is another possibility. We can define an additional communication channel with the simplest possible semantics: A signal that indicates how good the agent's current situation is. We will call this signal the *reward*. The agent's goal is then simply to maximise the amount of reward it receives, so in a sense its goal is fixed. This is not limiting though as we have not said anything about what causes different levels of reward to occur. In a complex setting the agent might be rewarded for winning a game or solving a difficult puzzle. From a broad perspective then, the goal is flexible. If the agent is to succeed in its environment, that is, receive a lot of reward, it must learn about the structure of the environment and in particular what it needs to do in order to get reward.

Not surprisingly, this is exactly the way in which we condition an animal to achieve a goal: by selectively rewarding certain behaviours. In a narrow sense the animal's goal is fixed, perhaps to get more treats to eat, but in a broader sense this may require doing a trick or solving a puzzle.

In our framework we will include the reward signal as a part of the perception generated by the environment. The perceptions also contain a non-reward part, which we will refer to as *observations*. This now gives us the complete system of interacting agent and environment in Figure 1. The goal, in the broad flexible sense, is implicitly defined by the environment as this is what defines when rewards are generated. Thus in this framework, to test an agent in any given way, it is sufficient to fully define the environment.

In artificial intelligence, this framework is used in the area of reinforcement learning [SB98]. By appropriately renaming things, it also describes the controller-plant framework used in control theory. It is a widely used and very general structure that can describe seemingly any kind of learning or control problem. The interesting point for us is that this type of framework follows naturally from our informal definition of intelligence. The only difficulty was how to deal with the notion of success, or profit. This requires the existence of some kind of objective or goal, and the most flexible and elegant way to bring this into our framework is by using a simple reward signal.

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender and Alexander Koller, Proceedings of the 58th Annual Meeting of the ACL, pp. 5185–5198 · July 2020

https://aclanthology.org/2020.acl-main.463/

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de

Abstract

The success of the large neural language models on many NLP tasks is exciting. However, we find that these successes sometimes lead to hype in which these models are being described as "understanding" language or capturing "meaning". In this position paper, we argue that a system trained only on form has *a priori* no way to learn meaning. In keeping with the ACL 2020 theme of "Taking Stock of Where We've Been and Where We're Going", we argue that a clear understanding of the distinction between form and meaning will help guide the field towards better science around natural language understanding.

1 Introduction

The current state of affairs in NLP is that the large neural language models (LMs), such as BERT (Devlin et al., 2019) or GPT-2 (Radford et al., 2019), are making great progress on a wide range of tasks, including those that are ostensibly meaning-sensitive. This has led to claims, in both academic and popular publications, that such models "understand" or "comprehend" natural language or learn its "meaning". From our perspective, these are overclaims caused by a misunderstanding of the relationship between linguistic form and meaning.

We argue that *the language modeling task, because it only uses form as training data, cannot in principle lead to learning of meaning*. We take the term *language model* to refer to any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not. We take (linguistic) *meaning* to be the relation between a linguistic form and communicative intent.

Our aim is to advocate for an alignment of claims and methodology: Human-analogous natural language understanding (NLU) is a grand challenge of artificial intelligence, which involves mastery of the structure and use of language and the ability to ground it in the world. While large neural LMs may well end up being important components of an eventual full-scale solution to human-analogous NLU, they are not nearly-there solutions to this grand challenge. We argue in this paper that genuine progress in our field — climbing the right hill, not just the hill on whose slope we currently sit — depends on maintaining clarity around big picture notions such as *meaning* and *understanding* in task design and reporting of experimental results.

After briefly reviewing the ways in which large LMs are spoken about and summarizing the recent flowering of "BERTology" papers (§2), we offer a working definition for "meaning" (§3) and a series of thought experiments illustrating the impossibility of learning meaning when it is not in the training signal (§4,5). We then consider the human language acquisition literature for insight into what information humans use to bootstrap language learning (§6) and the distributional semantics literature to discuss what is required to ground distributional models (§7). §8 presents reflections on how we look at progress and direct research effort in our field, and in §9, we address possible counterarguments to our main thesis.

2 Large LMs: Hype and analysis

Publications talking about the application of large LMs to meaning-sensitive tasks tend to describe the models with terminology that, if interpreted at face value, is misleading. Here is a selection from academically-oriented pieces (emphasis added):

(1) In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task. (Devlin et al., 2019)

(2) Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension** . . . (Ohsugi et al., 2019)

(3) The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demon-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5185–5198
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

strates their potential as unsupervised open-domain QA systems. (Petroni et al., 2019)

If the highlighted terms are meant to describe human-analogous understanding, comprehension, or recall of factual knowledge, then these are gross overclaims. If, instead, they are intended as technical terms, they should be explicitly defined.

One important consequence of imprudent use of terminology in our academic discourse is that it feeds AI hype in the popular press. As NLP gains public exposure and is more widely used in applied contexts, it is increasingly important that the actual capabilities of our systems be accurately represented. In some cases, NLP experts speaking with the media are being appropriately careful, as in these two quotes in the *New York Times*:[1]

(4) These systems are still a really long way from truly understanding running prose. (Gary Marcus)

(5) Though BERT passed the lab's common-sense test, machines are still a long way from an artificial version of a human's common sense. (Oren Etzioni)

However, there are plenty of instances where the popular press gets it wrong, such as (6) from the B2C website,[2] apparently based on the Google Blog post about BERT and search, which includes numerous statements like (7).[3]

(6) BERT is a system by which Google's algorithm uses pattern recognition to better **understand** how human beings communicate so that it can return more relevant results for users.

(7) Here are some of the examples that showed up our evaluation process that demonstrate BERTs ability to **understand** the intent behind your search.

In sum, it is not clear from our academic literature whether all authors are clear on the distinction between form and meaning, but it is clear that the way we speak about what neural LMs are doing is misleading to the public.

Part of the reason for this tendency to use imprecise language may well be that we do not yet fully understand what exactly it is about language that the large LMs come to implicitly represent. Their success, however, has sparked a subfield ('BERTology') that aims to answer this question. The methodology of probing tasks (e.g. Adi et al., 2017; Ettinger et al., 2018) has been used to show that large LMs learn at least some information about phenomena such as English subject-verb agreement (Goldberg, 2019; Jawahar et al., 2019), constituent types, dependency labels, NER, and (core) semantic role types (again, all in English) (Tenney et al., 2019).[4] Hewitt and Manning (2019) find information analogous to unlabeled dependency structures in the word vectors provided by ELMo and BERT (trained on English). And of course it is well established that vector-space representations of words pick up word classes, both syntactic (POS, e.g. Lin et al., 2015) and semantic (lexical similarity, e.g. Rubenstein and Goodenough, 1965; Mikolov et al., 2013).

Others have looked more closely at the success of the large LMs on apparently meaning sensitive tasks and found that in fact, far from doing the "reasoning" ostensibly required to complete the tasks, they were instead simply more effective at leveraging artifacts in the data than previous approaches. Niven and Kao (2019) find that BERT's unreasonably good performance on the English Argument Reasoning Comprehension Task (Habernal et al., 2018) falls back to chance if the dataset is modified by adding adversarial examples that just negate one piece of the original, thus mirroring the distribution of lexical cues for each label. Similarly, McCoy et al. (2019) find that BERT's performance on the English Multi-genre Natural Language Inference dataset (Williams et al., 2018) is predicated on its ability to leverage syntactic heuristics involving overlap (of full constituents, subsequences, or simply bags of words). In a dataset carefully designed to frustrate such heuristics, BERT's performance falls to significantly below chance.

In this brief overview of BERTology papers we have highlighted both the extent to which there is evidence that large LMs can learn aspects of linguistic formal structure (e.g. agreement, dependency structure), and how their apparent ability to "reason" is sometimes a mirage built on leveraging artifacts in the training data (i.e. form, not meaning). Our contribution is an argument on theoretical grounds that a system exposed only to form in its training cannot in principle learn meaning.

3 What is meaning?

We start by defining two key terms: We take *form* to be any observable realization of language: marks

[1] https://www.nytimes.com/2018/11/18/technology/artificial-intelligence-language.html, accessed 2019/12/04

[2] https://www.business2community.com/seo/what-to-do-about-bert-googles-recent-local-algorithm-update-02259261, accessed 2019/12/04

[3] https://www.blog.google/products/search/search-language-understanding-bert/, accessed 2019/12/04

[4] But see Warstadt et al.'s (2019) cautionary note about how the methodology used for probing can influence the results.

on a page, pixels or bytes in a digital representation of text, or movements of the articulators.[5] We take *meaning* to be the relation between the form and something external to language, in a sense that we will make precise below.

3.1 Meaning and communicative intent

When humans use language, we do so for a purpose: We do not talk for the joy of moving our articulators, but in order to achieve some *communicative intent*. There are many types of communicative intents: they may be to convey some information to the other person; or to ask them to do something; or simply to socialize. We take *meaning* to be the relation $M \subseteq E \times I$ which contains pairs (e, i) of natural language expressions e and the communicative intents i they can be used to evoke. Given this definition of meaning, we can now use *understand* to refer to the process of retrieving i given e.

Communicative intents are about something that is *outside of language*. When we say *Open the window!* or *When was Malala Yousafzai born?*, the communicative intent is grounded in the real world the speaker and listener inhabit together. Communicative intents can also be about abstract worlds, e.g. bank accounts, computer file systems, or a purely hypothetical world in the speaker's mind.

Linguists distinguish communicative intent from *conventional* (or *standing*) meaning (Quine, 1960; Grice, 1968). The conventional meaning of an expression (word, phrase, sentence) is what is constant across all of its possible contexts of use. Conventional meaning is an abstract object that represents the communicative potential of a form, given the linguistic system it is drawn from. Each linguistic system (say, English) provides a relation $C \subseteq E \times S$, which contains pairs (e, s) of expressions e and their conventional meanings s.[6] The field of linguistic semantics provides many competing theories of what conventional meanings s look like. For our purposes, we don't need to select among these theories; all we assume is that conventional meanings must have interpretations, such as a means of testing them for truth against a model of the world. Thus, like the meaning relation M, C connects language to objects outside of language.

[5] In spoken languages, the primary articulators are the components of the vocal tract. In signed languages, they are principally the hands and face.

[6] We abstract away here from the facts that linguistic systems C change over time and are only incompletely shared among different speakers. They are stable enough to function as rich signals to communicative intent.

Returning to the meaning relation M from above, it is best understood as mediated by the relation C of a linguistic system shared between two interlocutors. The speaker has a certain communicative intent i, and chooses an expression e with a standing meaning s which is fit to express i in the current communicative situation. Upon hearing e, the listener then reconstructs s and uses their own knowledge of the communicative situation and their hypotheses about the speaker's state of mind and intention in an attempt to deduce i.

This active participation of the listener is crucial to human communication (Reddy, 1979; Clark, 1996). For example, to make sense of (8) and (9) (from Clark, 1996, p.144), the listener has to calculate that *Napoleon* refers to a specific pose (hand inside coat flap) or that *China trip* refers to a person who has recently traveled to China.

(8) The photographer asked me to do a Napoleon for the camera.

(9) Never ask two China trips to the same party.

We humans are also very willing, as we will see in §4 below, to attribute communicative intent to a linguistic signal of a language we speak, even if the originator of the signal is not an entity that could have communicative intent.

To summarize, as we strive to understand how NLU tasks and system performance on those tasks relates to the bigger picture goals of building human-analogous natural language understanding systems, it is useful to distinguish cleanly between form, conventional meaning, and communicative intent. Furthermore, we should be careful not to confuse communicative intent with ground truth about the world, as speakers can of course be mistaken, be intentionally dissembling, etc.

We argue that a model of natural language that is trained purely on form will not learn meaning: if the training data is only form, there is not sufficient signal to learn the relation M between that form and the non-linguistic intent of human language users, nor C between form and the standing meaning the linguistic system assigns to each form.

3.2 Meaning and intelligence

Meaning and understanding have long been seen as key to intelligence. Turing (1950) argued that a machine can be said to "think" if a human judge cannot distinguish it from a human interlocutor after having an arbitrary written conversation with

each. However, humans are quick to attribute meaning and even intelligence to artificial agents, even when they know them to be artificial, as evidenced by the way people formed attachments to ELIZA (Weizenbaum, 1966; Block, 1981).

This means we must be extra careful in devising evaluations for machine understanding, as Searle (1980) elaborates with his Chinese Room experiment: he develops the metaphor of a "system" in which a person who does not speak Chinese answers Chinese questions by consulting a library of Chinese books according to predefined rules. From the outside, the system seems like it "understands" Chinese, although in reality no actual understanding happens anywhere inside the system.

Searle's thought experiment begins from the premise that it is possible to manipulate forms well enough to be indistinguishable from a system that understands the meaning of the forms, reasons about it, and responds appropriately. We observe that much recent work in NLP claims to be building systems where not only the runtime system but in fact also the process for building it only has access to form. But language is used for communication about the speakers' actual (physical, social, and mental) world, and so the reasoning behind producing meaningful responses must connect the meanings of perceived inputs to information about that world. This in turn means that for a human or a machine to learn a language, they must solve what Harnad (1990) calls the *symbol grounding problem*. Harnad encapsulates this by pointing to the impossibility for a non-speaker of Chinese to learn the meanings of Chinese words from Chinese dictionary definitions alone.

Our purpose here is to look more deeply into why meaning can't be learned from linguistic form alone, even in the context of modern hardware and techniques for scaling connectionist models to the point where they can take in vast amounts of data. We argue that, independently of whether passing the Turing test would mean a system is intelligent, a system that is trained only on form would fail a sufficiently sensitive test, because it lacks the ability to connect its utterances to the world.

4 The octopus test

In order to illustrate the challenges in attempting to learn meaning from form alone, we propose a concrete scenario. Say that A and B, both fluent speakers of English, are independently stranded on two uninhabited islands. They soon discover that previous visitors to these islands have left behind telegraphs and that they can communicate with each other via an underwater cable. A and B start happily typing messages to each other.

Meanwhile, O, a hyper-intelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations. O knows nothing about English initially, but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A's utterances. O also observes that certain words tend to occur in similar contexts, and perhaps learns to generalize across lexical patterns by hypothesizing that they can be used somewhat interchangeably. Nonetheless, O has never observed these objects, and thus would not be able to pick out the referent of a word when presented with a set of (physical) alternatives.

At some point, O starts feeling lonely. He cuts the underwater cable and inserts himself into the conversation, by pretending to be B and replying to A's messages. Can O successfully pose as B without making A suspicious? This constitutes a weak form of the Turing test (weak because A has no reason to suspect she is talking to a non-human); the interesting question is whether O fails it because he has not learned the meaning relation, having seen only the form of A and B's utterances.

The extent to which O can fool A depends on the task — that is, on what A is trying to talk about. A and B have spent a lot of time exchanging trivial notes about their daily lives to make the long island evenings more enjoyable. It seems possible that O would be able to produce new sentences of the kind B used to produce; essentially acting as a chatbot. This is because the utterances in such conversations have a primarily social function, and do not need to be grounded in the particulars of the interlocutors' actual physical situation nor anything else specific about the real world. It is sufficient to produce text that is internally coherent.

Now say that A has invented a new device, say a coconut catapult. She excitedly sends detailed instructions on building a coconut catapult to B, and asks about B's experiences and suggestions for improvements. Even if O had a way of constructing the catapult underwater, he does not know what words such as *rope* and *coconut* refer to, and thus can't physically reproduce the experiment. He can

only resort to earlier observations about how B responded to similarly worded utterances. Perhaps O can recognize utterances about *mangos* and *nails* as "similarly worded" because those words appeared in similar contexts as *coconut* and *rope*. So O decides to simply say "Cool idea, great job!", because B said that a lot when A talked about ropes and nails. It is absolutely conceivable that A accepts this reply as meaningful — but only because A does all the work in attributing meaning to O's response. It is not because O understood the meaning of A's instructions or even his own reply.

Finally, A faces an emergency. She is suddenly pursued by an angry bear. She grabs a couple of sticks and frantically asks B to come up with a way to construct a weapon to defend herself. Of course, O has no idea what A "means". Solving a task like this requires the ability to map accurately between words and real-world entities (as well as reasoning and creative thinking). It is at this point that O would fail the Turing test, if A hadn't been eaten by the bear before noticing the deception.[7]

Having only form available as training data, O did not learn meaning. The language exchanged by A and B is a projection of their communicative intents through the meaning relation into linguistic forms. Without access to a means of hypothesizing and testing the underlying communicative intents, reconstructing them from the forms alone is hopeless, and O's language use will eventually diverge from the language use of an agent who can ground their language in coherent communicative intents.

The thought experiment also illustrates our point from §3 about listeners' active role in communication. When O sent signals to A pretending to be B, he exploited statistical regularities in the form, i.e. the distribution of linguistic forms he observed. Whatever O learned is a reflection of A and B's communicative intents and the meaning relation. But reproducing this distribution is not sufficient for meaningful communication. O only fooled A into believing he was B because A was such an active listener: Because agents who produce English sentences usually have communicative intents, she assumes that O does too, and thus she builds the conventional meaning English associates with O's utterances. Because she assumes that O is B, she uses that conventional meaning together with her other guesses about B's state of mind and goals to attribute communicative intent. It is not that O's utterances make sense, but rather, that A can make sense of them.

[7] To see what a large LM might reply in this situation, we prompted the GPT-2 demo with *"Help! I'm being chased by a bear! All I have is these sticks. What should I do?"*, and GPT-2 to supplied *"You're not going to get away with this!"* (https://gpt2.apps.allenai.org/, accessed 2019/12/4). Following Radford et al.'s (2019) approach of giving explicit cues to encode the task, we also constructed a more elaborate prompt. The results, given in Appendix A, are highly entertaining but no more helpful to the hapless A.

5 More constrained thought experiments

The story of the octopus considers the problem of learning not only the full communicative system, including the relations M and C, but also the reasoning required to come up with answers that are both coherent and also helpful in the real world. Here, we provide two more constrained thought experiments, to focus more narrowly on the problem of learning the meaning relation, for both natural languages and programming languages.

Because programming languages are designed to be unambiguous and relatively insensitive to execution context, the distinction between standing and speaker meaning is less important than for natural languages. A Java program e, when compiled and executed on the Java Virtual Machine, can be interpreted as a function i which maps program inputs to program outputs. We take the meaning relation $J \subseteq E \times I$ of Java to contain all such pairs (e, i).

Java Imagine that we were to train an LM on all of the well-formed Java code published on Github. The input is only the code. It is not paired with bytecode, nor a compiler, nor sample inputs and outputs for any specific program. We can use any type of LM we like and train it for as long as we like. We then ask the model to execute a sample program, and expect correct program output.

English As as second example, imagine training an LM (again, of any type) on English text, again with no associated independent indications of speaker intent. The system is also given access to a very large collection of unlabeled photos, but without any connection between the text and the photos. For the text data, the training task is purely one of predicting form. For the image data, the training task could be anything, so long as it only involves the images. At test time, we present the model with inputs consisting of an utterance and a photograph, like *How many dogs in the picture are jumping?* or *Kim saw this picture and said "What a cute dog!" What is cute?* and the photos




Figure 1: Photo stimuli 1 (L) and 2 (R)

in Figure 1, where the appropriate answers are a number or a region of the photo, respectively.

Reflections In both cases, the tests are ridiculous. It seems patently unfair to ask the model to perform them, given what it was trained on. But that is precisely the point we are trying to make: a system that has learned the meaning (semantics) of a programming language knows how to execute code in that language. And a system that has learned the meaning of a human language can do things like answer questions posed in the language about things in the world (or in this case, in pictures).

In other words, what's interesting here is not that the tasks are impossible, but rather what makes them impossible: what's missing from the training data. The form of Java programs, to a system that has not observed the inputs and outputs of these programs, does not include information on how to execute them. Similarly, the form of English sentences, to a system that has not had a chance to acquire the meaning relation C of English, and in the absence of any signal of communicative intent, does not include any information about what language-external entities the speaker might be referring to. Accordingly, a system trained only on the form of Java or English has no way learn their respective meaning relations.

6 Human language acquisition

One common reason for believing LMs *might* be learning meaning is the claim that human children can acquire language just by listening to it. This is not supported by scholarly work on language acquisition: rather, we find that human language learning is not only grounded in the physical world around us, but also in interaction with other people in that world. Kids won't pick up a language from passive exposure such as TV or radio: Snow et al. (1976) note in passing that Dutch-speaking kids who watch German TV shows by choice nonetheless don't learn German. Kuhl (2007) shows experimentally that English-learning infants can learn Mandarin phonemic distinctions from brief interactions with a Mandarin-speaking experimenter but not from exposure to Mandarin TV or radio.

Baldwin (1995) and others argue that what is critical for language learning is not just interaction but actually joint attention, i.e. situations where the child and a caregiver are both attending to the same thing and both aware of this fact. This theoretical perspective is substantiated with experimental results showing that toddlers (observed at 15 and 21 months) whose caregivers "follow into" their attention and provide labels for the object of joint attention more have larger vocabularies (Tomasello and Farrar, 1986); that toddlers (18–20 months old) don't pick up labels uttered by someone behind a screen, but do pick up labels uttered by someone performing joint attention with them (Baldwin, 1995); and that at around 10–11 months of age babies pay attention to whether a person's eyes are open or not in terms of whether to follow their gaze, and the degree to which infants in fact follow gaze at 10–11 months while vocalizing themselves predicts vocabulary comprehension 7–8 months later (Brooks and Meltzoff, 2005).[8]

In summary, the process of acquiring a linguistic system, like human communication generally, relies on joint attention and intersubjectivity: the ability to be aware of what another human is attending to and guess what they are intending to communicate. Human children do not learn meaning from form alone and we should not expect machines to do so either.

7 Distributional semantics

Distributional semanticists have long been aware that grounding distributional representations in the real world is challenging. The lexical similarity relations learned by distributional models trained on text don't in themselves connect any of those words to the world (Herbelot, 2013; Baroni et al., 2014; Erk, 2016; Emerson, 2020), and the distributions of words may not match the distribution of things in the world (consider *four-legged dogs*).

One approach to providing grounding is to train distributional models on corpora augmented with perceptual data, such as photos (Hossain et al., 2019) or other modalities (Kiela and Clark, 2015; Kiela et al., 2015). Another is to look to interaction data, e.g. a dialogue corpus with success annotations, including low-level success signals such as

[8] These three studies do not name the language that the children were learning. It appears to have been English.

emotional stress (McDuff and Kapoor, 2019) or eye gaze (Koller et al., 2012), which contains a signal about the felicitous uses of forms. The idea that as the learner gets access to more and more information in addition to the text itself, it can learn more and more facets of meaning is worked out in detail by Bisk et al. (2020). We agree that this is an exciting avenue of research.

From this literature we can see that the slogan "meaning is use" (often attributed to Wittgenstein, 1953), refers not to "use" as "distribution in a text corpus" but rather that language is *used* in the real world to convey communicative intents to real people. Speakers distill their past experience of language use into what we call "meaning" here, and produce new attempts at using language based on this; this attempt is successful if the listener correctly deduces the speaker's communicative intent. Thus, standing meanings evolve over time as speakers can different experiences (e.g. McConnell-Ginet, 1984), and a reflection of such change can be observed in their changing textual distribution (e.g. Herbelot et al., 2012; Hamilton et al., 2016).

8 On climbing the right hills

What about systems which are trained on a task that is not language modeling — say, semantic parsing, or reading comprehension tests — and that use word embeddings from BERT or some other large LM as one component? Numerous papers over the past couple of years have shown that using such pretrained embeddings can boost the accuracy of the downstream system drastically, even for tasks that are clearly related to meaning.

Our arguments do not apply to such scenarios: reading comprehension datasets include information which goes beyond just form, in that they specify semantic relations between pieces of text, and thus a sufficiently sophisticated neural model *might* learn some aspects of meaning when trained on such datasets. It also is conceivable that whatever information a pretrained LM captures might help the downstream task in learning meaning, without being meaning itself.

Recent research suggests that it is wise to interpret such findings with caution. As noted in §2, both McCoy et al. (2019) and Niven and Kao (2019) found that BERT picked up idiosyncratic patterns in the data for their tasks, and not "meaning". Beyond such diagnostic research on why large pretrained LMs boost such tasks so much, we think there is a more fundamental question to be asked here: Are we climbing the right hill?

8.1 Top-down and bottom-up theory-building

There are two different perspectives from which one can look at the progress of a field. Under a *bottom-up* perspective, the efforts of a scientific community are driven by identifying specific research challenges. A scientific result counts as a success if it solves such a specific challenge, at least partially. As long as such successes are frequent and satisfying, there is a general atmosphere of sustained progress. By contrast, under a *top-down* perspective, the focus is on the remote end goal of offering a complete, unified theory for the entire field. This view invites anxiety about the fact that we have not yet fully explained all phenomena and raises the question of whether all of our bottom-up progress leads us in the right direction.

There is no doubt that NLP is currently in the process of rapid hill-climbing. Every year, states of the art across many NLP tasks are being improved significantly — often through the use of better pretrained LMs — and tasks that seemed impossible not long ago are already old news. Thus, everything is going great when we take the bottom-up view. But from a top-down perspective, the question is whether the hill we are climbing so rapidly is the *right* hill. How do we know that incremental progress on today's tasks will take us to our end goal, whether that is "General Linguistic Intelligence" (Yogatama et al., 2019) or a system that passes the Turing test or a system that captures the meaning of English, Arapaho, Thai, or Hausa to a linguist's satisfaction?

It is instructive to look at the past to appreciate this question. Computational linguistics has gone through many fashion cycles over the course of its history. Grammar- and knowledge-based methods gave way to statistical methods, and today most research incorporates neural methods. Researchers of each generation felt like they were solving relevant problems and making constant progress, from a bottom-up perspective. However, eventually serious shortcomings of each paradigm emerged, which could not be tackled satisfactorily with the methods of the day, and these methods were seen as obsolete. This negative judgment — we were climbing a hill, but not the right hill — can only be made from a top-down perspective. We have discussed the question of what is required to

learn meaning in an attempt to bring the top-down perspective into clearer focus.

8.2 Hillclimbing diagnostics

We can only definitively tell if we've been climbing the right hill in hindsight, but we propose some best practices for less error-prone mountaineering:

First, above all, cultivate humility towards language and ask top-down questions. Neural methods are not the first bottom-up success in NLP; they will probably not be the last.

Second, be aware of the limitations of tasks: Artificial tasks like bAbI (Weston et al., 2016) can help get a field of research off the ground, but there is no reason to assume that the distribution of language in the test data remotely resembles the distribution of real natural language; thus evaluation results on such tasks must be interpreted very carefully. Similar points can be made about crowdsourced NLI datasets such as SQuAD (Rajpurkar et al., 2016) or SNLI (Bowman et al., 2015), which do not represent questions that any particular person really wanted to ask about a text, but the somewhat unnatural communicative situation of crowdsourcing work. If a system does better on such a task than the inter-annotator agreement,[9] the task probably has statistical artifacts that do not represent meaning. In the vision community, Barbu et al. (2019) offer a novel dataset which explicitly tries to achieve a more realistic distribution of task data; it would be interesting to explore similar ideas for language.

Third, value and support the work of carefully creating new tasks (see also Heinzerling, 2019). For example, the DROP reading comprehension benchmark (Dua et al., 2019) seeks to create more stringent tests of understanding by creating questions that require the system to integrate information from different parts of a paragraph via simple arithmetic or similar operations.[10]

Fourth, evaluate models of meaning across tasks. (Standing) meaning is task-independent, so a system that captures meaning should do well on multiple tasks. Efforts like SuperGLUE (Wang et al., 2019) seem like a good step in this direction.

Finally, perform thorough analysis of both errors and successes. As McCoy et al. (2019) and Niven and Kao (2019) have shown, systems that find success with large pretrained LMs do not necessarily do so because the LMs have learned "meaning". Analyses which start from an attitude of healthy skepticism ("too good to be true") and probing tasks which try to identify what the model actually learned can be good ways to find out whether the system performs well for the right reasons.

[9] https://rajpurkar.github.io/SQuAD-explorer/

[10] See Appendix B for an exploration of what GPT-2 does with arithmetic.

9 Some possible counterarguments

In discussing the main thesis of this paper with various colleagues over the past 18 months, we have observed recurring counterarguments. In this section, we address those counterarguments, plus a few more that might arise.

"But 'meaning' doesn't mean what you say it means." Defining "meaning" is notoriously hard. For the purposes of this paper, we chose a working definition which is as general as we could make it, capturing the crucial point that meaning is based on the link between linguistic form and something that is not language. "Meaning" cannot simply be the relation between form and some kind of "deep syntax", e.g. semantic dependency graphs (Oepen et al., 2015); like syntax, such representations could perhaps be learned from form alone (He et al., 2018; Hewitt and Manning, 2019). Equating these with meaning ignores a core function of language, which is to convey communicative intents.

"But meaning could be learned from . . . ". As we discussed in §7, if form is augmented with grounding data of some kind, then meaning can conceivably be learned to the extent that the communicative intent is represented in that data.

In addition, certain tasks are designed in a way that specific forms are declared as representing certain semantic relations of interest. Examples of this include NLI datasets (Dagan et al., 2006; Rajpurkar et al., 2016; Ostermann et al., 2019) which pair input/output tuples of linguistic forms with an explicit semantic relation (e.g. text + hypothesis + "entailed"). Similarly, control codes, or tokens like *tl;dr*, have been used to prompt large LMs to perform summarization and other tasks (Radford et al., 2019; Keskar et al., 2019). Here forms are explicitly declared at test time to represent certain semantic relations, which together with the distributional similarity between e.g. *tl;dr* and other phrases such as *in summary*, may be enough to bootstrap a successful neural summarizer. Depending on one's perspective, one may argue that such a system has learned to reliably find instances of the relation without understanding the text; or that

explicitly declaring cues like *entailed* or *tl;dr* as representing certain semantic relations provides a training signal that goes beyond pure form.

Analogously, it has been pointed out to us that the sum of all Java code on Github (cf. § 5) contains unit tests, which specify input-output pairs for Java code. Thus a learner could have access to a weak form of interaction data, from which the meaning of Java could conceivably be learned. This is true, but requires a learner which has been equipped by its human developer with the ability to identify and interpret unit tests. This learner thus has access to partial grounding in addition to the form.

"But there is *so much* form out there – surely that is enough." We have argued for the general principle that learning meaning requires more than form. How much form can be observed is not relevant to our point; the octopus can observe A and B for as long as he wants, and the quantity of training data in §5 is not limited.

But given lots of form, could O perhaps learn to keep producing seemingly meaningful responses to A's utterances without learning meaning? The problem is that people constantly generate new communicative intents to talk about their constantly evolving inner and outer worlds, and thus O would need to memorize infinitely many stimulus-response pairs. Such an approach may be an avenue towards high scores in evaluations where perfection is not expected anyway; but it is probably not an avenue towards human-analogous NLU.

"But aren't neural representations meaning too?" The internal representations of a neural network have been found to capture certain aspects of meaning, such as semantic similarity (Mikolov et al., 2013; Clark, 2015). As we argued in §4, semantic similarity is only a weak reflection of actual meaning. Neural representations neither qualify as standing meanings (s), lacking interpretations, nor as communicative intents (i), being insufficient to e.g. correctly build a coconut catapult.

An interesting recent development is the emergence of models for unsupervised machine translation trained only with a language modeling objective on monolingual corpora for the two languages (Lample et al., 2018). If such models were to reach the accuracy of supervised translation models, this would seem contradict our conclusion that meaning cannot be learned from form. A perhaps surprising consequence of our argument would then be that accurate machine translation does not actually require a system to understand the meaning of the source or target language sentence.

"But BERT improves performance on meaning-related tasks, so it must have learned something about meaning." It has probably learned *something* about meaning, in the same sense that syntax captures something about meaning and semantic similarity captures something about meaning: a potentially useful, but incomplete, reflection of the actual meaning. McCoy et al. (2019) and Niven and Kao (2019) provide cautionary tales about overestimating what that "something" is purely based on evaluation results on existing tasks. What exactly BERT and its relatives learn about meaning is a very interesting question, and we look forward to further findings from the field of BERTology.

10 Conclusion

In this paper, we have argued that in contrast to some current hype, meaning cannot be learned from form alone. This means that even large language models such as BERT do not learn "meaning"; they learn some reflection of meaning into the linguistic form which is very useful in applications.

We have offered some thoughts on how to maintain a healthy, but not exaggerated, optimism with respect to research that builds upon these LMs. In particular, this paper can be seen as a call for precise language use when talking about the success of current models and for humility in dealing with natural language. With this we hope to encourage a top-down perspective on our field which we think will help us select the right hill to climb towards human-analogous NLU.

Acknowledgments. This paper benefitted from many inspiring and often spirited discussions. Without implying any agreement with the contents as presented, we thank Sam Bowman, Vera Demberg, Lucia Donatelli, Jason Eisner, Jonas Groschwitz, Kristen Howell, Angie McMillan-Major, Joakim Nivre, Stephan Oepen, Ellie Pavlick, Benjamin Roth, Dan Roth, Asad Sayeed, Hinrich Schütze, Nina Tahmasebi, and Olga Zamaraeva. This paper originated in a Twitter mega-thread that was neatly summarized by Thomas Wolf (2018). We also thank the ACL reviewers and the participants of the Toulouse Workshop on Formal and Distributional Semantics (2015) and *SEM 2016 for their insightful and constructive thoughts.

The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents

Nick Bostrom, Minds and Machines 22(2): 71–85 · 2012
Extract: §1–2.2
https://nickbostrom.com/superintelligentwill.pdf

THE SUPERINTELLIGENT WILL: MOTIVATION AND INSTRUMENTAL RATIONALITY IN ADVANCED ARTIFICIAL AGENTS

(2012) Nick Bostrom
Future of Humanity Institute
Faculty of Philosophy & Oxford Martin School
Oxford University
www.nickbostrom.com

[*Minds and Machines*, Vol. 22, Iss. 2, May 2012] [translation: Portuguese]

ABSTRACT
This paper discusses the relation between intelligence and motivation in artificial agents, developing and briefly arguing for two theses. The first, the *orthogonality thesis*, holds (with some caveats) that *intelligence* and *final goals* (purposes) are orthogonal axes along which possible artificial intellects can freely vary—more or less any level of intelligence could be combined with more or less any final goal. The second, the *instrumental convergence thesis*, holds that as long as they possess a sufficient level of intelligence, agents having any of a wide range of final goals will pursue similar intermediary goals because they have instrumental reasons to do so. In combination, the two theses help us understand the possible range of behavior of superintelligent agents, and they point to some potential dangers in building such an agent.

KEYWORDS: superintelligence, artificial intelligence, AI, goal, instrumental reason, intelligent agent

1. The orthogonality of motivation and intelligence

1.1 Avoiding anthropomorphism

If we imagine a space in which all possible minds can be represented, we must imagine all *human* minds as constituting a small and fairly tight cluster within that space. The personality differences between Hannah Arendt and Benny Hill might seem vast to us, but this is because the scale bar in our intuitive judgment is calibrated on the existing human distribution. In the wider space of all logical possibilities, these two personalities are close neighbors. In terms of neural architecture, at least, Ms. Arendt and Mr. Hill are nearly identical. Imagine their brains laying side by side in quiet repose. The differences would appear minor and you would quite readily recognize them as two of a kind; you might even be unable to tell which brain was whose. If you studied the morphology of the two brains more closely under a microscope, the

impression of fundamental similarity would only be strengthened: you would then see the same lamellar organization of the cortex, made up of the same types of neuron, soaking in the same bath of neurotransmitter molecules.[1]

It is well known that naïve observers often anthropomorphize the capabilities of simpler insensate systems. We might say, for example, "This vending machine is taking a long time to think about my hot chocolate." This might lead one either to underestimate the cognitive complexity of capabilities which come naturally to human beings, such as motor control and sensory perception, or, alternatively, to ascribe significant degrees of mindfulness and intelligence to very dumb systems, such as chatterboxes like Weizenbaum's ELIZA (Weizenbaum 1976). In a similar manner, there is a common tendency to anthropomorphize the *motivations* of intelligent systems in which there is really no ground for expecting human-like drives and passions ("My car really didn't want to start this morning"). Eliezer Yudkowsky gives a nice illustration of this phenomenon:

> Back in the era of pulp science fiction, magazine covers occasionally depicted a sentient monstrous alien—colloquially known as a bug-eyed monster (BEM)—carrying off an attractive human female in a torn dress. It would seem the artist believed that a non-humanoid alien, with a wholly different evolutionary history, would sexually desire human females … Probably the artist did not ask whether a giant bug *perceives* human females as attractive. Rather, a human female in a torn dress *is sexy*—inherently so, as an intrinsic property. They who made this mistake did not think about the insectoid's mind: they focused on the woman's torn dress. If the dress were not torn, the woman would be less sexy; the BEM does not enter into it. (Yudkowsky 2008)

An artificial intelligence can be far less human-like in its motivations than a space alien. The extraterrestrial (let us assume) is a biological creature who has arisen through a process of evolution and may therefore be expected to have the kinds of motivation typical of evolved creatures. For example, it would not be hugely surprising to find that some random intelligent alien would have motives related to the attaining or avoiding of food, air, temperature, energy expenditure, the threat or occurrence of bodily injury, disease, predators, reproduction, or protection of offspring. A member of an intelligent social species might also have motivations related to cooperation and competition: like us, it might show in-group loyalty, a resentment of free-riders, perhaps even a concern with reputation and appearance.

By contrast, an artificial mind need not care intrinsically about any of those things, not even to the slightest degree. One can easily conceive of an artificial intelligence whose sole fundamental goal is to count the grains of sand on Boracay, or to calculate decimal places of pi indefinitely, or to maximize the total number of paperclips in its future lightcone. In fact, it would be easier to create an AI with simple goals like these, than to build one that has a human-like set of values and dispositions.

[1] This is of course not to deny that differences that appear small visually can be functionally profound.

1.2 The orthogonality thesis

For our purposes, "intelligence" will be roughly taken to correspond to the capacity for instrumental reasoning (more on this later). Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal. Intelligence and motivation can in this sense be thought of as a pair of orthogonal axes on a graph whose points represent intelligent agents of different paired specifications. Each point in the graph represents a logically possible artificial agent, modulo some weak constraints—for instance, it might be impossible for a very unintelligent system to have very complex motivations, since complex motivations would place significant demands on memory. Furthermore, in order for an agent to "have" a set of motivations, this set may need to be functionally integrated with the agent's decision-processes, which again would place demands on processing power and perhaps on intelligence. For minds that can modify themselves, there may also be dynamical constraints; for instance, an intelligent mind with an urgent desire to be stupid might not remain intelligent for very long. But these qualifications should not obscure the main idea, which we can express as follows:

> *The Orthogonality Thesis*
> Intelligence and final goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could in principle be combined with more or less any final goal.

A comparison may be made here with the Humean theory of motivation. David Hume thought that beliefs alone (say, about what is a good thing to do) cannot motivate action: some desire is required.[2] This would support the orthogonality thesis by undercutting one possible objection to it, namely, that sufficient intelligence might entail the acquisition of certain beliefs, and that these beliefs would necessarily produce certain motivations. Not so, according to David Hume: belief and motive are separate.

Although the orthogonality thesis can draw support from the Humean theory of motivation, it does not presuppose it. In particular, one need not maintain that beliefs alone can *never* motivate action. It would suffice to assume, for example, that an agent—be it ever so intelligent—can be motivated to pursue any course of action *if* the agent happens to have certain standing desires of some sufficient, overriding strength. Another way in which the orthogonality thesis could be true even if the Humean theory of motivation is false is if arbitrarily high intelligence does not entail the acquisition of any such beliefs as are (putatively) motivating on their own. A third way in which it might be possible for the orthogonality thesis to be true even if the Humean theory were false is if it is possible to build a cognitive system (or more neutrally, an "optimization process") with arbitrarily high intelligence but with constitution so alien as to contain no clear functional analogues to what in humans we call

[2] For some recent attempts to defend the Humean theory of motivation, see Smith (1987), Lewis (1988), and Sinhababu (2009).

"beliefs" and "desires". This would be the case if such a system could be constructed in a way that would make it motivated to pursue any given final goal.

The orthogonality thesis, as formulated here, makes a claim about the relationship between motivation and *intelligence*, rather than between motivation and *rationality* (or motivation and *reason*). This is because some philosophers use the word "rationality" to connote a "normatively thicker" concept than we seek to connote here with the word "intelligence". For instance, in *Reasons and Persons* Derek Parfit argues that certain basic preferences would be irrational, such as that of an otherwise normal agent who has "Future-Tuesday-Indifference":

> A certain hedonist cares greatly about the quality of his future experiences. With one exception, he cares equally about all the parts of his future. The exception is that he has Future-Tuesday-Indifference. Throughout every Tuesday he cares in the normal way about what is happening to him. But he never cares about possible pains or pleasures on a future Tuesday... This indifference is a bare fact. When he is planning his future, it is simply true that he always prefers the prospect of great suffering on a Tuesday to the mildest pain on any other day. (Parfit 1984)[3]

Thus, the agent is now indifferent to his own future suffering if and only if it occurs on a future Tuesday. For our purposes, we need take no stand on whether Parfit is right that this is irrational, so long as we grant that it is not necessarily unintelligent. By "intelligence" here we mean something like *instrumental rationality*—skill at prediction, planning, and means-ends reasoning in general. Parfit's imaginary Future-Tuesday-Indifferent agent could have impeccable instrumental rationality, and therefore great intelligence, even if he falls short on some kind of sensitivity to "objective reason" that might be required of a fully rational agent. Consequently, this kind of example does not undermine the orthogonality thesis.

In a similar vein, even if there are objective moral facts that any fully rational agent would comprehend, and even if these moral facts are somehow intrinsically motivating (such that anybody who fully comprehends them is necessarily motivated to act in accordance with them) this need not undermine the orthogonality thesis. The thesis could still be true if an agent could have impeccable *instrumental* rationality even whilst lacking some other faculty constitutive of rationality proper, or some faculty required for the full comprehension of the objective moral facts. (An agent could also be extremely intelligent, even superintelligent, without having full instrumental rationality in every domain.)

One reason for focusing on intelligence, that is, on instrumental rationality, is that this is the most relevant concept if we are trying to figure out what different kinds of systems would do. Normative questions, such as whether their behavior would count as being prudentially rational or morally justifiable, can be important in various ways. However, such questions should not blind us to the possibility of cognitive systems that fail to satisfy substantial

[3] See also Parfit (2011).

normative criteria but which are nevertheless very powerful and able to exert strong influence on the world.[4]

1.3 Predicting superintelligence motivation and behavior

The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip-maximizing. This holds even (indeed *especially*) for artificial agents that are extremely intelligent or superintelligent. Yet it does not follow from the orthogonality thesis that it is impossible to make predictions about what particular agents will do. Predictability is important if one seeks to design a system to achieve particular outcomes, and the issue becomes more important the more powerful the artificial agent in question is. Superintelligent agents could be *extremely* powerful, so it is important to develop a way of analyzing and predicting their behavior. Yet despite the independence of intelligence and final goals implied by the orthogonality thesis, the problem of predicting an agent's behavior need not be intractable—not even with regard to hypothetical superintelligent agents whose cognitive complexity and performance characteristics might render them in certain respects opaque to human analysis.

There are at least three directions from which one can approach the problem of predicting superintelligent motivation:

(1) *Predictability through design competence*. If we can suppose that the designers of a superintelligent agent can successfully engineer the goal system of the agent so that it stably pursues a particular goal set by the programmers, then one prediction we can make is that the agent will pursue that goal. The more intelligent the agent is, the greater the cognitive resourcefulness it will have to pursue that goal. So even before an agent has been created we might be able to predict something about its behavior, if we know something about who will build it and what goals they will want it to have.

(2) *Predictability through inheritance.* If a digital intelligence is created directly from a human template (as would be the case in a high-fidelity whole brain emulation), then the digital intelligence might inherit the motivations of the human template.[5] The agent might retain some of these motivations even if its cognitive capacities are subsequently enhanced to make it superintelligent. This kind of inference requires caution. The agent's goals and values could easily become corrupted in the uploading process or during its subsequent operation and enhancement, depending on how the procedure is implemented.

[4] The orthogonality thesis implies that most any combination of final goal and intelligence level is logically possible; it does *not* imply that it would be practically easy to endow a superintelligent agent with some arbitrary or human-respecting final goal—even if we knew how to construct the intelligence part. For some preliminary notes on the value-loading problem, see, e.g., Dewey (2011) and Yudkowsky (2011).
[5] See Sandberg & Bostrom (2008).

(3) *Predictability through convergent instrumental reasons*. Even without detailed knowledge of an agent's final goals, we may be able to infer something about its more immediate objectives by considering the *instrumental* reasons that would arise for any of a wide range of possible final goals in a wide range of situations. This way of predicting becomes more useful the greater the intelligence of the agent, because a more intelligent agent is more likely to recognize the true instrumental reasons for its actions, and so act in ways that make it more likely to achieve its goals.

The next section explores this third way of predictability and develops an "instrumental convergence thesis" which complements the orthogonality thesis.

2. Instrumental convergence

According to the orthogonality thesis, artificial intelligent agents may have an enormous range of possible final goals. Nevertheless, according to what we may term the "instrumental convergence" thesis, there are some *instrumental* goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal. We can formulate this thesis as follows:

> *The Instrumental Convergence Thesis*
> Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by many intelligent agents.

In the following we will consider several categories where such convergent instrumental values may be found.[6] The likelihood that an agent will recognize the instrumental values it confronts increases (ceteris paribus) with the agent's intelligence. We will therefore focus mainly on the case of a hypothetical superintelligent agent whose instrumental reasoning capacities far

[6] Stephen Omohundro has written two pioneering papers on this topic (Omohundro 2008a, 2008b). Omohundro argues that all advanced AI systems are likely to exhibit a number of "basic drives", by which he means "tendencies which will be present unless explicitly counteracted." The term "AI drive" has the advantage of being short and evocative, but it has the disadvantage of suggesting that the instrumental goals to which it refers influence the AI's decision-making in the same way as psychological drives influence human decision-making, i.e. via a kind of phenomenological tug on our ego which our willpower may occasionally succeed in resisting. That connotation is unhelpful. One would not normally say that a typical human being has a "drive" to fill out their tax return, even though filing taxes may be a fairly convergent instrumental goal for humans in contemporary societies (a goal whose realization averts trouble that would prevent us from realizing many of our final goals). Our treatment here also differs from that of Omohundro in some other more substantial ways, although the underlying idea is the same. (See also Chalmers (2010) and Omohundro (2012).

exceed those of any human. We will also comment on how the instrumental convergence thesis applies to the case of human beings, as this gives us occasion to elaborate some essential qualifications concerning how the instrumental convergence thesis should be interpreted and applied. Where there are convergent instrumental values, we may be able to predict some aspects of a superintelligence's behavior even if we know virtually nothing about that superintelligence's final goals.

2.1 Self-preservation

Suppose that an agent has some final goal that extends some way into the future. There are many scenarios in which the agent, if it is still around in the future, is then able to perform actions that increase the probability of achieving the goal. This creates an instrumental reason for the agent to try to be around in the future—to help achieve its present future-oriented goal.

Agents with human-like motivational structures often seem to place some *final* value on their own survival. This is not a necessary feature of artificial agents: some may be designed to place no final value whatever on their own survival. Nevertheless, even agents that do not care intrinsically about their own survival would, under a fairly wide range of conditions, care instrumentally to some degree about their own survival in order to accomplish the final goals they do value.

2.2 Goal-content integrity

An agent is more likely to act in the future to maximize the realization of its present final goals if it still has those goals in the future. This gives the agent a present instrumental reason to prevent alterations of its final goals. (This argument applies only to final goals. In order to attain its final goals, an intelligent agent will of course routinely want to change its subgoals in light of new information and insight.)

Goal-content integrity for final goals is in a sense even more fundamental than survival as a convergent instrumental motivation. Among humans, the opposite may seem to be the case, but that is because survival is usually part of our final goals. For software agents, which can easily switch bodies or create exact duplicates of themselves, preservation of self as a particular implementation or a particular physical object need not be an important instrumental value. Advanced software agents might also be able to swap memories, download skills, and radically modify their cognitive architecture and personalities. A population of such agents might operate more like a "functional soup" than a society composed of distinct semi-permanent persons.[7] For some purposes, processes in such a system might be better individuated as *teleological threads*, based on their final values, rather than on the basis of bodies, personalities, memories, or abilities. In such scenarios, goal-continuity might be said to *constitute* a key aspect of survival.

Even so, there are situations in which an agent may intentionally change its own final goals. Such situations can arise when any of the following factors is significant:

[7] See Chislenko (1997).

- *Social signaling*. When others can perceive an agent's goals and use that information to infer instrumentally relevant dispositions or other correlated attributes, it can be in the agent's interest to modify its goals to make whatever desired impression. For example, an agent might miss out on beneficial deals if potential partners cannot trust it to fulfill its side of the bargain. In order to make credible commitments, an agent might therefore wish to adopt as a final goal the honoring of its earlier commitments, and to allow others to verify that it has indeed adopted this goal. Agents that could flexibly and transparently modify their own goals could use this ability to enforce deals among one another.[8]
- *Social preferences*. Others may also have preferences about an agent's goals. The agent could then have reason to modify its goals, either to satisfy or to frustrate those preferences.
- *Preferences concerning own goal content*. An agent might have some final goal concerned with the agent's own goal content. For example, the agent might have a final goal to become the type of agent that is motivated by certain values, such as compassion.
- *Storage costs*. If the cost of storing or processing some part of an agent's utility function is large compared to the chance that a situation will arise in which applying that part of the utility function will make a difference, then the agent has an instrumental reason to simplify its goal content, and it may trash that part of the utility function.[9] [10]

We humans often seem happy to let our final goals and values drift. This might often be because we do not know precisely what they are. We obviously want our *beliefs* about our final goals and values to be able to change in light of continuing self-discovery or changing self-presentation needs. However, there are cases in which we willingly change the goals and values themselves, not just our beliefs or interpretations of them. For example, somebody deciding to have a child might predict that they will come to value the child for its own sake, even though at the time of the decision they may not particularly value their future child or even like children in general.

[8] See also Shulman (2010).

[9] An agent might also change its goal *representation* if it changes its ontology, in order to transpose its old representation into the new ontology. Cf. de Blanc (2011).

[10] Another type of factor that might make an *evidential decision theorist* undertake various actions, including changing its final goals, is the evidential import of deciding to do so. For example, an agent that follows evidential decision theory might believe that there exist other agents like it in the universe, and that its own actions will provide some evidence about how those other agents will act. The agent might therefore choose to adopt a final goal that is altruistic towards those other evidentially-linked agents, on grounds that this will give the agent evidence that those other agents will have chosen to act in like manner. An equivalent outcome might be obtained, however, without changing one's final goals, by choosing in each instant to act *as if* one had those final goals.

Humans are complicated, and many factors might be at play in a situation like this.[11] For instance, one might have a final value that involves becoming the kind of person who cares about some other individual for his or her own sake (here one places a final value on having a certain final value). Alternatively, one might have a final value that involves having certain experiences and occupying a certain social role; and becoming a parent—and undergoing an associated goal shift—might be a necessary part of that. Human goals can also have inconsistent content, goal content; and so some people might want to modify some of their final goals to reduce the inconsistencies.

[11] An extensive psychological literature explores adaptive preference formation. See, e.g., Forgas *et al.* (2009).

The AI Cargo Cult: The Myth of a Superhuman AI

Kevin Kelly, Backchannel · 25 April 2017
https://kk.org/wp-content/uploads/2017/05/Myth_Superhuman_Backchannel.pdf

I've heard that in the future computerized AIs will become so much smarter than us that they will take all our jobs and resources, and humans will go extinct. Is this true?

That's the most common question I get whenever I give a talk about AI. The questioners are earnest; their worry stems in part from some experts who are asking themselves the same thing. These folks are some of the smartest people alive today, such as Stephen Hawking, Elon Musk, Max Tegmark, Sam Harris, and Bill Gates, and they believe this scenario very likely could be true. Recently at a conference convened to discuss these AI issues, a panel of nine of the most informed gurus on AI all agreed this superhuman intelligence was inevitable and not far away.

Yet buried in this scenario of a takeover of superhuman artificial intelligence are five assumptions which, when examined closely, are not based on any evidence. These claims might be true in the future, but there is no evidence to date to support them. The assumptions behind a superhuman intelligence arising soon are:

1. Artificial intelligence is already getting smarter than us, at an exponential rate.
2. We'll make AIs into a general purpose intelligence, like our own.
3. We can make human intelligence in silicon.
4. Intelligence can be expanded without limit.
5. Once we have exploding superintelligence it can solve most of our problems.

In contradistinction to this orthodoxy, I find the following five heresies to have more evidence to support them.

1. Intelligence is not a single dimension, so "smarter than humans" is a meaningless concept.
2. Humans do not have general purpose minds, and neither will AIs.
3. Emulation of human thinking in other media will be constrained by cost.
4. Dimensions of intelligence are not infinite.
5. Intelligences are only one factor in progress.

If the expectation of a superhuman AI takeover is built on five key assumptions that have no basis in evidence, then this idea is more akin to a religious belief — a myth. In the following paragraphs I expand my evidence for each of these five counter-assumptions, and make the case that, indeed, a superhuman AI is a kind of myth.

1.

The most common misconception about artificial intelligence begins with the common misconception about natural intelligence. This misconception is that intelligence is a single dimension. Most technical people tend to graph intelligence the way Nick Bostrom does in his book, Superintelligence — as a literal, single-dimension, linear graph of increasing amplitude. At one end is the low intelligence of, say, a small animal; at the other end is the high intelligence, of, say, a genius—almost as if intelligence were a sound level in decibels. Of course, it is then very easy to imagine the extension so that the loudness of intelligence continues to grow, eventually to exceed our own high intelligence and become a super-loud intelligence — a roar! — way beyond us, and maybe even off the chart.

This model is topologically equivalent to a ladder, so that each rung of intelligence is a step higher than the one before. Inferior animals are situated on lower rungs below us, while higher-level intelligence AIs will inevitably overstep us onto higher rungs. Time scales of when it happens are not important; what is important is the ranking—the metric of increasing intelligence.

The problem with this model is that it is mythical, like the ladder of evolution. The pre-Darwinian view of the natural world supposed a ladder of being, with inferior animals residing on rungs below human. Even post-Darwin, a very common notion is the "ladder" of evolution, with fish evolving into reptiles, then up a step into mammals, up into primates, into humans, each one a little more evolved (and of course smarter) than the one before it. So the ladder of intelligence parallels the ladder of existence. But both of these models supply a thoroughly unscientific view.

A more accurate chart of the natural evolution of species is a disk radiating outward, like this one (above) first devised by David Hillis at the University of Texas and based on DNA. This deep genealogy mandala begins in the middle with the most primeval life forms, and then branches outward in time. Time moves outward so that the most recent species of life living on the planet today form the perimeter of the circumference of this circle. This picture emphasizes a fundamental fact of evolution that is hard to appreciate: Every species alive today is equally evolved. Humans exist on this outer ring alongside cockroaches, clams, ferns, foxes, and bacteria. Every one of these species has

undergone an unbroken chain of three billion years of successful reproduction, which means that bacteria and cockroaches today are as highly evolved as humans. There is no ladder.

Likewise, there is no ladder of intelligence. Intelligence is not a single dimension. It is a complex of many types and modes of cognition, each one a continuum. Let's take the very simple task of measuring animal intelligence. If intelligence were a single dimension we should be able to arrange the intelligences of a parrot, a dolphin, a horse, a squirrel, an octopus, a blue whale, a cat, and a gorilla in the correct ascending order in a line. We currently have no scientific evidence of such a line. One reason might be that there is no difference between animal intelligences, but we don't see that either. Zoology is full of remarkable differences in how animals think. But maybe they all have the same relative "general intelligence?" It could be, but we have no measurement, no single metric for that intelligence. Instead we have many different metrics for many different types of cognition.

Instead of a single decibel line, a more accurate model for intelligence is to chart its possibility space, like the above rendering of possible forms created by an algorithm written by Richard Dawkins. Intelligence is a combinatorial continuum. Multiple nodes, each node a continuum, create complexes of high diversity in high dimensions. Some intelligences may be very complex, with many sub-nodes of thinking. Others may be simpler but more extreme, off in a corner of the space. These complexes we call intelligences might be thought of Generated via Emergent Mind as symphonies comprising many types of instruments. They vary not only in loudness, but also in pitch, melody, color, tempo, and so on. We could think of them as ecosystem. And in that sense, the different component nodes of thinking are co-dependent, and co-created.

Human minds are societies of minds, in the words of Marvin Minsky. We run on ecosystems of thinking. We contain multiple species of cognition that do many types of thinking: deduction, induction, symbolic reasoning, emotional intelligence, spacial logic, short-term memory, and long-term memory. The entire nervous system in our gut is also a type of brain with its own mode of cognition. We don't really think with just our brain; rather, we think with our whole bodies.

These suites of cognition vary between individuals and between species. A squirrel can remember the exact location of several thousand acorns for years, a feat that blows human minds away. So in that one type of cognition, squirrels exceed humans. That superpower is bundled with some other modes that are dim compared to ours in order to produce a squirrel mind. There are many other specific feats of cognition in the animal kingdom that are superior to humans, again bundled into different systems.

Likewise in AI. Artificial minds already exceed humans in certain dimensions. Your calculator is a genius in math; Google's memory is already beyond our own in a certain dimension. We are engineering AIs to excel in specific modes. Some of these modes are things we can do, but they can do better, such as probability or math. Others are type of thinking we can't do at all — memorize every single word on six billion web pages, a feat any search engine can do. In the future, we will invent whole new modes of cognition that don't exist in us and don't exist anywhere in biology. When we invented artificial flying we were inspired by biological modes of flying, primarily flapping wings. But the flying we invented — propellers bolted to a wide fixed wing — was a new mode of flying unknown in our biological world. It is alien flying. Similarly, we will invent whole new modes of thinking that do not exist in nature. In many cases they will be new, narrow, "small," specific modes for specific jobs — perhaps a type of reasoning only useful in statistics and probability.

In other cases the new mind will be complex types of cognition that we can use to solve problems our intelligence alone cannot. Some of the hardest problems in business and science may require a two-step solution. Step one is: Invent a new mode of thought to work with our minds. Step two: Combine to solve the problem. Because we are solving problems we could not solve before, we want to call this cognition "smarter" than us, but really it is different than us. It's the differences in thinking that are the main benefits of AI. I think a useful model of AI is to think of it as alien intelligence (or artificial aliens). Its alienness will be its chief asset.

At the same time we will integrate these various modes of cognition into more complicated, complex societies of mind. Some of these complexes will be more complex than us, and because they will be able to solve problems we can't, some will want to call them superhuman. But we don't call Google a superhuman AI even though its memory is beyond us, because there are many things we can do better than it. These complexes of artificial intelligences will for sure be able to exceed us in many dimensions, but no one entity will do all we do better. It's similar to the physical powers of humans. The industrial revolution is 200 years old, and while all machines as a class can beat the physical achievements of an individual human (speed of running, weight lifting, precision cutting, etc.), there is no one machine that can beat an average human in everything he or she does.

Even as the society of minds in an AI become more complex, that complexity is hard to measure scientifically at the moment. We don't have good operational metrics of complexity that could determine whether a cucumber is more complex than a Boeing 747, or the ways their complexity might differ. That is one of the reasons why we don't have

good metrics for smartness as well. It will become very difficult to ascertain whether mind A is more complex than mind B, and for the same reason to declare whether mind A is smarter than mind B. We will soon arrive at the obvious realization that “smartness” is not a single dimension, and that what we really care about are the many other ways in which intelligence operates — all the other nodes of cognition we have not yet discovered.

2.

The second misconception about human intelligence is our belief that we have a general purpose intelligence. This repeated belief influences a commonly stated goal of AI researchers to create an artificial general purpose intelligence (AGI). However, if we view intelligence as providing a large possibility space, there is no general purpose state. Human intelligence is not in some central position, with other specialized intelligence revolving around it. Rather, human intelligence is a very, very specific type of intelligence that has evolved over many millions of years to enable our species to survive on this planet. Mapped in the space of all possible intelligences, a human-type of intelligence will be stuck in the corner somewhere, just as our world is stuck at the edge of vast galaxy.

We can certainly imagine, and even invent, a Swiss-army knife type of thinking. It kind of does a bunch of things okay, but none of them very well. AIs will follow the same engineering maxim that all things made or born must follow: You cannot optimize every dimension. You can only have tradeoffs. You can’t have a general multi-purpose unit outperform specialized functions. A big “do everything” mind can’t do everything as well as those things done by specialized agents. Because we believe our human minds are general purpose, we tend to believe that cognition does not follow the engineer’s tradeoff, that it will be possible to build an intelligence that maximizes all modes of thinking. But I see no evidence of that. We simply haven’t invented enough varieties of minds to see the full space (and so far we have tended to dismiss animal minds as a singular type with variable amplitude on a single dimension.)

3.

Part of this belief in maximum general-purpose thinking comes from the concept of universal computation. Formally described as the Church-Turing hypothesis in 1950, this conjecture states that all computation that meets a certain threshold is equivalent. Therefore there is a universal core to all computation, whether it occurs in one machine with many fast parts, or slow parts, or even if it occurs in a biological brain, it is the same logical process. Which means that you should be able to emulate any computational process (thinking) in any machine that can do “universal” computation. Singularitans rely on this principle for their expectation that we will be able to engineer silicon brains to hold human minds, and that we can make artificial minds that think like humans, only much smarter. We should be skeptical of this hope because it relies on a misunderstanding of the Church-Turing hypothesis.

The starting point of the theory is: “Given infinite tape [memory] and time, all computation is equivalent.” The problem is that in reality, no computer has infinite memory or time. When you are operating in the real world, real time makes a huge difference, often a life-or-death difference. Yes, all thinking is equivalent if you ignore time. Yes, you can emulate human-type thinking in any matrix you want, as long as you ignore time or the real-life constraints of storage and memory. However, if you incorporate time, then you have to restate the principal in a significant way: Two computing systems operating on vastly different platforms won’t be equivalent in real time. That can be restated again as: The only way to have equivalent modes of thinking is to run them on equivalent substrates. The physical matter you run your computation on — particularly as it gets more complex — greatly influences the type of cognition that can be done well in real time.

I will extend that further to claim that the only way to get a very human-like thought process is to run the computation on very human-like wet tissue. That also means that very big, complex artificial intelligences run on dry silicon will produce big, complex, unhuman-like minds. If it would be possible to build artificial wet brains using human-like grown neurons, my prediction is that their thought will be more similar to ours. The benefits of such a wet brain are proportional to how similar we make the substrate. The costs of creating wetware is huge and the closer that tissue is to human brain tissue, the more cost-efficient it is to just make a human. After all, making a human is something we can do in nine months.

Furthermore, as mentioned above, we think with our whole bodies, not just with our minds. We have plenty of data showing how our gut’s nervous system guides our “rational” decision-making processes, and can predict and learn. The more we model the entire human body system, the closer we get to replicating it. An intelligence running on a very different body (in dry silicon instead of wet carbon) would think differently.

I don’t see that as a bug but rather as a feature. As I argue in point 2, thinking differently from humans is AI’s chief asset. This is yet another reason why calling it “smarter than humans” is misleading and misguided.

4.

At the core of the notion of a superhuman intelligence — particularly the view that this intelligence will keep improving itself — is the essential belief that intelligence has an infinite scale. I find no evidence for this. Again, mistaking intelligence as a single dimension helps this belief, but we should understand it as a belief. There is no other physical dimension in the universe that is infinite, as far as science knows so far. Temperature is not infinite — there is finite cold and finite heat. There is finite space and time. Finite speed. Perhaps the mathematical number line is infinite, but all other physical attributes are finite. It stands to reason that reason itself is finite, and not infinite. So the question is, where is the limit of intelligence? We tend to believe that the limit is way beyond us, way "above" us, as we are "above" an ant. Setting aside the recurring problem of a single dimension, what evidence do we have that the limit is not us? Why can't we be at the maximum? Or maybe the limits are only a short distance away from us? Why do we believe that intelligence is something that can continue to expand forever?

A much better way to think about this is to see our intelligence as one of a million types of possible intelligences. So while each dimension of cognition and computation has a limit, if there are hundreds of dimensions, then there are uncountable varieties of mind — none of them infinite in any dimension. As we build or encounter these uncountable varieties of mind we might naturally think of some of them as exceeding us. In my recent book The Inevitable, I sketched out some of that variety of minds that were superior to us in some way. Here is an incomplete list:

- A mind like a human mind, just faster in answering (the easiest AI mind to imagine).
- A very slow mind, composed primarily of vast storage and memory.
- A global supermind composed of millions of individual dumb minds in concert.
- A hive mind made of many very smart minds, but unaware it/they are a hive.
- A borg supermind composed of many smart minds that are very aware they form a unity.
- A mind trained and dedicated to enhancing your personal mind, but useless to anyone else.
- A mind capable of imagining a greater mind, but incapable of making it.
- A mind capable of creating a greater mind, but not self-aware enough to imagine it.
- A mind capable of successfully making a greater mind, once.
- A mind capable of creating a greater mind that can create a yet greater mind, etc.

Some folks today may want to call each of these entities a superhuman AI, but the sheer variety and alienness of these minds will steer us to new vocabularies and insights about intelligence and smartness.

Second, believers of Superhuman AI assume intelligence will increase exponentially (in some unidentified single metric), probably because they also assume it is already expanding exponentially. However, there is zero evidence so far that intelligence — no matter how you measure it — is increasing exponentially. By exponential growth I mean that artificial intelligence doubles in power on some regular interval. Where is that evidence? Nowhere I can find. If there is none now, why do we assume it will happen soon? The only thing expanding on an exponential curve are the inputs in AI, the resources devoted to producing the smartness or intelligences. But the output performance is not on a Moore's law rise. AIs are not getting twice as smart every 3 years, or even every 10 years.

I asked a lot of AI experts for evidence that intelligence performance is on an exponential gain, but all agreed we don't have metrics for intelligence, and besides, it wasn't working that way. When I asked Ray Kurzweil, the exponential wizard himself, where the evidence for exponential AI was, he wrote to me that AI does not increase explosively but rather by levels. He said: "It takes an exponential improvement both in computation and algorithmic complexity to add each additional level to the hierarchy…. So we can expect to add levels linearly because it requires exponentially more complexity to add each additional layer, and we are indeed making exponential progress in our ability to do this. We are not that many levels away from being comparable to what the neocortex can do, so my 2029 date continues to look comfortable to me."

What Ray seems to be saying is that it is not that the power of artificial intelligence is exploding exponentially, but that the effort to produce it is exploding exponentially, while the output is merely raising a level at a time. This is almost the opposite of the assumption that intelligence is exploding. This could change at some time in the future, but artificial intelligence is clearly not increasing exponentially now.

Therefore when we imagine an "intelligence explosion," we should imagine it not as a cascading boom but rather as a scattering exfoliation of new varieties. A Cambrian explosion rather than a nuclear explosion. The results of accelerating technology will most likely not be super-human, but extra-human. Outside of our experience, but not necessarily "above" it.

5.

Another unchallenged belief of a super AI takeover, with little evidence, is that a super, near-infinite intelligence can quickly solve our major unsolved problems.

Many proponents of an explosion of intelligence expect it will produce an explosion of progress. I call this mythical belief “thinkism.” It’s the fallacy that future levels of progress are only hindered by a lack of thinking power, or intelligence. (I might also note that the belief that thinking is the magic super ingredient to a cure-all is held by a lot of guys who like to think.)

Let’s take curing cancer or prolonging longevity. These are problems that thinking alone cannot solve. No amount of thinkism will discover how the cell ages, or how telomeres fall off. No intelligence, no matter how super duper, can figure out how the human body works simply by reading all the known scientific literature in the world today and then contemplating it. No super AI can simply think about all the current and past nuclear fission experiments and then come up with working nuclear fusion in a day. A lot more than just thinking is needed to move between not knowing how things work and knowing how they work. There are tons of experiments in the real world, each of which yields tons and tons of contradictory data, requiring further experiments that will be required to form the correct working hypothesis. Thinking about the potential data will not yield the correct data.

Thinking (intelligence) is only part of science; maybe even a small part. As one example, we don’t have enough proper data to come close to solving the death problem. In the case of working with living organisms, most of these experiments take calendar time. The slow metabolism of a cell cannot be sped up. They take years, or months, or at least days, to get results. If we want to know what happens to subatomic particles, we can’t just think about them. We have to build very large, very complex, very tricky physical structures to find out. Even if the smartest physicists were 1,000 times smarter than they are now, without a Collider, they will know nothing new.

There is no doubt that a super AI can accelerate the process of science. We can make computer simulations of atoms or cells and we can keep speeding them up by many factors, but two issues limit the usefulness of simulations in obtaining instant progress. First, simulations and models can only be faster than their subjects because they leave something out. That is the nature of a model or simulation. Also worth noting: The testing, vetting and proving of those models also has to take place in calendar time to match the rate of their subjects. The testing of ground truth can’t be sped up.

These simplified versions in a simulation are useful in winnowing down the most promising paths, so they can accelerate progress. But there is no excess in reality; everything real makes a difference to some extent; that is one definition of reality. As models and simulations are beefed up with more and more detail, they come up against the limit that reality runs faster than a 100 percent complete simulation of it. That is another definition of reality: the fastest possible version of all the details and degrees of freedom present. If you were able to model all the molecules in a cell and all the cells in a human body, this simulation would not run as fast as a human body. No matter how much you thought about it, you still need to take time to do experiments, whether in real systems or in simulated systems.

To be useful, artificial intelligences have to be embodied in the world, and that world will often set their pace of innovations. Without conducting experiments, building prototypes, having failures, and engaging in reality, an intelligence can have thoughts but not results. There won’t be instant discoveries the minute, hour, day or year a so-called “smarter-than-human” AI appears. Certainly the rate of discovery will be significantly accelerated by AI advances, in part because alien-ish AI will ask questions no human would ask, but even a vastly powerful (compared to us) intelligence doesn’t mean instant progress. Problems need far more than just intelligence to be solved.

Not only are cancer and longevity problems that intelligence alone can’t solve, so is intelligence itself. The common trope among Singularitans is that once you make an AI “smarter than humans” then all of sudden it thinks hard and invents an AI “smarter than itself,” which thinks harder and invents one yet smarter, until it explodes in power, almost becoming godlike. We have no evidence that merely thinking about intelligence is enough to create new levels of intelligence. This kind of thinkism is a belief. We have a lot of evidence that in addition to great quantities of intelligence we need experiments, data, trial and error, weird lines of questioning, and all kinds of things beyond smartness to invent new kinds of successful minds. The Seven Steps Toward Making Humans Obsolete Or, how we stopped fighting and learned to love our robot replacements. Excerpted from The Inevitable by Kevin Kelly.

I’d conclude by saying that I could be wrong about these claims. We are in the early days. We might discover a universal metric for intelligence; we might discover it is infinite in all directions. Because we know so little about what intelligence is (let alone consciousness), the possibility of some kind of AI singularity is greater than zero. I think all the evidence suggests that such a scenario is highly unlikely, but it is greater than zero.

So while I disagree on its probability, I am in agreement with the wider aims of OpenAI and the smart people who worry about a superhuman AI — that we should engineer friendly AIs and figure out how to instill self-replicating values that match ours. Though I think a superhuman AI is a remote possible existential threat (and worthy of considering), I think its unlikeliness (based on the evidence we have so far) should not be the guide for our science, policies, and development. An asteroid strike on the Earth would be catastrophic. Its probability is greater than zero

(and so we should support the B612 Foundation), but we shouldn't let the possibility of an asteroid strike govern our efforts in, say, climate change, or space travel, or even city planning.

Likewise, the evidence so far suggests AIs most likely won't be superhuman but will be many hundreds of extra-human new species of thinking, most different from humans, none that will be general purpose, and none that will be an instant god solving major problems in a flash. Instead there will be a galaxy of finite intelligences, working in unfamiliar dimensions, exceeding our thinking in many of them, working together with us in time to solve existing problems and create new problems.

I understand the beautiful attraction of a superhuman AI god. It's like a new Superman. But like Superman, it is a mythical figure.

Somewhere in the universe a Superman might exist, but he is very unlikely. However myths can be useful, and once invented they won't go away. The idea of a Superman will never die. The idea of a superhuman AI Singularity, now that it has been birthed, will never go away either. But we should recognize that it is a religious idea at this moment and not a scientific one. If we inspect the evidence we have so far about intelligence, artificial and natural, we can only conclude that our speculations about a mythical superhuman AI god are just that: myths.

Many isolated islands in Micronesia made their first contact with the outside world during World War II. Alien gods flew over their skies in noisy birds, dropped food and goods on their islands, and never returned. Religious cults sprang up on the islands praying to the gods to return and drop more cargo. Even now, fifty years later, many still wait for the cargo to return. It is possible that superhuman AI could turn out to be another cargo cult. A century from now, people may look back to this time as the moment when believers began to expect a superhuman AI to appear at any moment and deliver them goods of unimaginable value. Decade after decade they wait for the superhuman AI to appear, certain that it must arrive soon with its cargo.

Yet non-superhuman artificial intelligence is already here, for real. We keep redefining it, increasing its difficulty, which imprisons it in the future, but in the wider sense of alien intelligences — of a continuous spectrum of various smartness, intelligences, cognition, reasonings, learning, and consciousness — AI is already pervasive on this planet and will continue to spread, deepen, diversify, and amplify. No invention before will match its power to change our world, and by century's end AI will touch and remake everything in our lives. Still the myth of a superhuman AI, poised to either gift us super-abundance or smite us into super-slavery (or both), will probably remain alive—a possibility too mythical to dismiss.

Machines of Loving Grace: How AI Could Transform the World for the Better

Dario Amodei · October 2024

Extract: Basic assumptions and framework

https://www.darioamodei.com/essay/machines-of-loving-grace



Basic assumptions and framework

To make this whole essay more precise and grounded, it's helpful to specify clearly what we mean by powerful AI (i.e. the threshold at which the 5-10 year clock starts counting), as well as laying out a framework for thinking about the effects of such AI once it's present.

What powerful AI (I dislike the term AGI) will look like, and when (or if) it will arrive, is a huge topic in itself. It's one I've discussed publicly and could write a completely separate essay on (I probably will at some point). Obviously, many people are skeptical that powerful AI will be built soon and some are skeptical that it will ever be built at all. I think it could come as early as 2026, though there are also ways it could take much longer. But for the purposes of this essay, I'd like to put these issues aside, assume it will come reasonably soon, and focus on what happens in the 5-10 years after that. I also want to assume a definition of what such a system will look like, what its capabilities are and how it interacts, even though there is room for disagreement on this.

By powerful AI, I have in mind an AI model—likely similar to today's LLMs in form, though it might be based on a different architecture, might involve several interacting models, and might be trained differently—with the following properties:

- In terms of pure intelligence, it is smarter than a Nobel Prize winner across most relevant fields – biology, programming, math, engineering, writing, etc. This means it can prove unsolved mathematical theorems, write extremely good novels, write difficult codebases from scratch, etc.
- In addition to just being a "smart thing you talk to", it has all the "interfaces" available to a human working virtually, including text, audio, video, mouse and keyboard control, and internet access. It can engage in any actions, communications, or remote operations enabled by this interface, including taking actions on the internet, taking or giving directions to humans, ordering materials, directing experiments, watching videos, making videos, and so on. It does all of these tasks with, again, a skill exceeding that of the most capable humans in the world.
- It does not just passively answer questions; instead, it can be given tasks that take hours, days, or weeks to complete, and then goes off and does those tasks autonomously, in the way a smart employee would, asking for clarification as necessary.
- It does not have a physical embodiment (other than living on a computer screen), but it can control existing physical tools, robots, or laboratory equipment through a computer; in theory it could even design robots or equipment for itself to use.
- The resources used to train the model can be repurposed to run millions of instances of it (this matches projected cluster sizes by ~2027), and the model can absorb information and generate actions at roughly 10x-100x human speed. It may however be limited by the response time of the physical world or of software it interacts with.

- Each of these million copies can act independently on unrelated tasks, or if needed can all work together in the same way humans would collaborate, perhaps with different subpopulations fine-tuned to be especially good at particular tasks.

We could summarize this as a “country of geniuses in a datacenter”.

Clearly such an entity would be capable of solving very difficult problems, very fast, but it is not trivial to figure out how fast. Two “extreme” positions both seem false to me. First, you might think that the world would be instantly transformed on the scale of seconds or days (“the Singularity”), as superior intelligence builds on itself and solves every possible scientific, engineering, and operational task almost immediately. The problem with this is that there are real physical and practical limits, for example around building hardware or conducting biological experiments. Even a new country of geniuses would hit up against these limits. Intelligence may be very powerful, but it isn’t magic fairy dust.

Second, and conversely, you might believe that technological progress is saturated or rate-limited by real world data or by social factors, and that better-than-human intelligence will add very little. This seems equally implausible to me —I can think of hundreds of scientific or even social problems where a large group of really smart people would drastically speed up progress, especially if they aren’t limited to analysis and can make things happen in the real world (which our postulated country of geniuses can, including by directing or assisting teams of humans).

I think the truth is likely to be some messy admixture of these two extreme pictures, something that varies by task and field and is very subtle in its details. I believe we need new frameworks to think about these details in a productive way.

Other Links

What happens when our computers get smarter than we are?
Nick Bostrom, TED2015 (talk and transcript) · March 2015
https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are

How Long Before Superintelligence?
Nick Bostrom, International Journal of Futures Studies 2 · 1998
https://nickbostrom.com/papers/how-long-before-superintelligence/

Superintelligence: Paths, Dangers, Strategies
Nick Bostrom, Oxford University Press · 2014
https://nickbostrom.com/superintelligence

The implausibility of intelligence explosion
François Chollet, Medium · 27 November 2017
https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec

On the Measure of Intelligence
François Chollet, arXiv:1911.01547 · November 2019
https://arxiv.org/abs/1911.01547

A Collection of Definitions of Intelligence
Shane Legg and Marcus Hutter, arXiv:0706.3639 · June 2007
https://arxiv.org/abs/0706.3639

The Singularity: A Philosophical Analysis
David J. Chalmers, Journal of Consciousness Studies 17: 7–65 · 2010
https://consc.net/papers/singularity.pdf

Minds, Brains, and Programs
John R. Searle, Behavioral and Brain Sciences 3(3): 417–457 · 1980
https://doi.org/10.1017/S0140525X00005756

Why AI is Harder Than We Think
Melanie Mitchell, arXiv:2104.12871 · April 2021
https://arxiv.org/abs/2104.12871

The Gentle Singularity
Sam Altman · 10 June 2025
https://blog.samaltman.com/the-gentle-singularity

AI 2027
Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland and Romeo Dean · 3 April 2025
https://ai-2027.com/