The BigMother Manifesto: A Roadmap to Provably Maximally-Aligned Maximally-Superintelligent AGI (Part 1)

[unfinished draft]

The BigMother Manifesto: A Roadmap to Provably Maximally-Aligned Maximally-Superintelligent AGI (Part 1)

Aaron Turner AARON.TURNER@BIGMOTHER.AI
BigMother.AI CIC
Cambridge, UK

Timestamp: 2024-04-24 18:42:18Z

Abstract

Whoever owns human-level AGI (Artificial General Intelligence) will own the global means of production for all goods and services. Superintelligent AGI has been conservatively estimated to have a net present value of $13.5 quadrillion. Accordingly, the major profit-motivated AI labs (and their associated sovereign states) are currently engaged in an AGI arms race, each in pursuit of their own short-term self-interest, seemingly oblivious to the long-term best interest of the human species as a whole.

The currently dominant opinion among AI and AI safety researchers seems to be that Large Language Models (LLMs), built using the Transformer neural network model (or similar), massively scaled, and aligned with human preferences via RLHF (Reinforcement Learning through Human Feedback) and other methods, represent the most promising path to AGI and beyond, with the median estimate for when human-level AGI will arrive ranging from 2026 to 2031, with superintelligent AGI arriving ~2 years later.

At the same time, many AI researchers hold variously negative opinions about LLMs, including that "we have no idea how they work", that they are merely "stochastic parrots" capable of at most weak reasoning over shallow world models, that LLM hallucinations are inevitable, that the celebrated scaling laws will never get us to AGI, and that reliable LLM alignment is impossible. In a recent survey, the median machine learning researcher appeared to put a 5-10% chance on the extinction risk from misaligned AGI.

We propose an alternative to the de facto LLM-based short-term-self-interest-driven approach. Rather than engage in a race, over the next 10-20 years, towards an AGI future that is likely to be, at best, sub-optimal for all mankind *for all eternity* (due to the "trap-door" nature of superintelligence), and, at worst, catastrophic, we propose spending 50-100 years doing it properly, in the best interest of all mankind, in order to achieve an endgame that is (as close as possible to) maximally-beneficent and minimally-maleficent for all mankind, while at the same time using the additional breathing space to mitigate to the maximum extent possible the inevitable pain of such a profound transition.

Our overall approach is to try to imagine the *ideal endgame*, and to work backwards from there in order to make it (or something close to it) actually happen. This is largely equivalent to imagining the ideal (or "Gold Standard") superintelligent AGI, and then working backwards to actually build it. To this end, we seek to design, develop, and deploy a *provably maximally-aligned maximally-superintelligent* AGI (called BigMother/BigMom) that is publicly owned by all mankind (via the United Nations), and whose operation benefits all mankind, without favouring any subset thereof (such as the citizens of any particular country or countries, or the shareholders of any particular company or companies).

In this paper, we describe the BigMother cognitive architecture and associated BigMother project in detail. Together, these define an AGI research agenda for the next 50-100 years.

Keywords: Artificial General Intelligence, cognitive architecture, superintelligence, alignment

For Sir Clive Sinclair (30 July 1940 – 16 September 2021)

Contents

1. Overview

This is a very long (OK, book-length!) paper. Key words and phrases are highlighted in **bold** in order to facilitate **skimming** on first reading. The structure of the paper may be summarised as follows:

- In Section 2, we:
  - explain why the current paper is the first in a **sequence of BigMother papers**
  - introduce a number of basic **concepts and definitions** pertaining to AGI[1]
  - explore the concept of **consciousness** in relation to AGI
  - describe the **specific problem** that we seek to address
  - describe our **proposed approach** to the problem that we have identified.
- In Section 3, we describe a **cognitive architecture** for a proposed **Gold Standard AGI**
  - this section includes a tentative solution to the **AI alignment** problem.
- In Section 4, we describe a **construction sequence** for the proposed Gold Standard AGI
  - any AGI so constructed will be **well-founded** by virtue of its method of construction.
- In Section 5, we explore the problem of **AGI governance**.
- In Section 6, we propose a **coordinated collaboration** of international experts to:
  - **safely build and deploy** the proposed Gold Standard AGI
  - participate in the global **AGI governance** process.
- In Section 7, we offer our **conclusions**.
- In Section 8, we thank our **informal reviewers** for their contributions.

[1] **Artificial General Intelligence**

2. Introduction

This paper describes work conducted by the author from 1985 to 2024, but not previously reported[2].

2.1 It takes a village

Artificial General Intelligence (AGI) [Turing (1950); Goertzel and Pennachin (2007); Goertzel and Wang (2007); Wang and Goertzel (2012)] — the central subject of this paper — is **complicated**. Thus (i) the opportunities for **miscommunication** by an author, and **misunderstanding** by a reader, are endless, and (ii) **it takes a village to build an AGI**, and a particularly large and varied village to build a **superintelligent** AGI [Good (1966); Bostrom (2014); Yampolskiy (2016)].

On the one hand, it is extremely desirable, when embarking upon a technical project, to have a description of that project that is **accessible** to technical and non-technical project contributors alike — i.e. the entire village — in order that everyone involved has a **baseline conceptual understanding** of what it is that they're actually doing. (In particular, the dominant factor determining the accessibility of a document will be how mathematical it is.) On the other hand, more **advanced readers** will expect a **much deeper technical exposition**. One document cannot satisfy both audiences!

Accordingly, we envisage a sequence of BigMother papers as described in Table 1:

Table 1: The envisaged sequence of BigMother papers

Part	Accessible?	Description
1	yes	assumes only basic high-school mathematics (plus some determination!)
≥ 2	no	expands upon its predecessor; may contain more advanced mathematics

Although quite long (currently 51 pages), Part 1 is nevertheless designed to be an **easy read**. Given the complexities of AGI, **Part 1 is a good starting point even for more advanced readers**.

2.2 Basic concepts and definitions pertaining to AGI

2.2.1 WHAT IS INTELLIGENCE?

It's not really possible to answer questions about intelligence, Artificial Intelligence (AI), or AGI without first considering what these things mean. Despite prior attempts at providing formal definitions [e.g. Barr, Feigenbaum, and Cohen (1982); Gottfredson (1997); Nilsson (2010); Russell and Norvig (2021); Wang (2019); Monett, Lewis, and Thórisson (2020)], **the concept of intelligence** remains elusive, i.e. "intelligence" means different things to different people. In the absence of any widely accepted definitions, we will attempt to define these concepts relative to our specific purposes.

2.2.2 PHYSICAL SYSTEMS

Assuming that the physical universe exists (an assumption to which we will return later), any **physical system** (i.e. any part of the physical universe) may potentially be intelligent (or not) relative to some definition. Physical systems include rocks, cats, humans, organisations, computers, and robots.

[2] *** funding acknowledgement goes here

2.2.3 EXTRINSIC INTELLIGENCE

From the perspective of an **external observer**, a physical system is either a **black box** or a **white box**. If it's a white box then we can see inside it (i.e. we can see *at least some* of its internal structure); if it's a black box then we cannot. In the case of a black box, the only information we have pertaining to its intelligence is its **externally observable behaviour**, i.e. its pattern of interaction with its physical environment. As a first approximation, a physical system possesses **extrinsic intelligence** if an external observer **deems it to be behaving intelligently** (relative to the observer's own intuitive, and thus subjective, understanding of "intelligence") on the basis of its externally observable behaviour[3].

2.2.4 ANALOGOUS INTELLIGENCE

On what other basis might an external observer deem a physical system to be behaving intelligently (or otherwise)? Let's imagine that an external observer (i) has already concluded that physical system $\mathcal{A}$ is intelligent, and (ii) has determined, after a period of observation, that physical system $\mathcal{B}$**'s pattern of external behaviour** is **analogous** to physical system $\mathcal{A}$**'s pattern of external behaviour** (in other words, if certain aspects are ignored, and others are retained, then $\mathcal{A}$ and $\mathcal{B}$ may be regarded as behaving "equivalently"). If conditions (i) and (ii) are satisfied then $\mathcal{A}$ is deemed to be intelligent (by i), $\mathcal{A}$ and $\mathcal{B}$ are deemed to be "equivalent" (by ii), and therefore $\mathcal{B}$ may be deemed to be intelligent.

Note that the precise nature of the analogy applied at step (ii) — specifically, which aspects of externally observable behaviour are ignored, and which are retained — is of paramount importance. Different analogies (focusing on different aspects) may lead to diametrically opposite conclusions.

2.2.5 HUMAN-LIKE INTELLIGENCE

Given that *intelligence* is deemed to be the quality that most distinguishes humans [Harari (2011)] from other species, **it is natural to use humans as the reference intelligence** $\mathcal{A}$ against which physical system $\mathcal{B}$ is compared in order to determine (by analogy) whether or not the latter is intelligent.

As already alluded, the immediate problem that then arises is the exact nature of the analogy to be used when forming the comparison, i.e. which aspects of externally observable human behaviour are deemed relevant to intelligence, and which are not. For example, is an ability to converse in natural language a requirement for human-like intelligence? Or an ability to draw pictures and diagrams? Or an ability to compose plays, poetry, and music? The exact choice of analogy is highly subjective, and consequently so is the definition of "human-like intelligence" that results from using this approach.

In order to explore *intelligence* systematically, **we need a more objective definition** than this.

2.2.6 PROBLEM-SOLVING

If we're **engineers** (meaning that we build stuff) and all we really care about is **utility** (i.e. that the systems that we build are practically useful in some way) then we might choose to equate **intelligence** with **problem solving** (i.e. the ability to solve problems). From this perspective, whenever we're *thinking*, we're solving some problem (internally, in our minds), and a *mind* is "something that thinks" or "something that solves problems". Whenever an intelligent system (i.e. a mind) solves a problem, some utility (real-world value) is generated. This definition seems to work reasonably well. For our present purposes, therefore, let's define **intelligence** as **problem solving**, and see where it leads us[4].

[3] without further qualification, this definition is highly subjective — different observers may make conflicting judgements

[4] this definition, while more objective, is deliberately broad (even a simple pocket calculator will qualify as intelligent)

2.2.7 ARTIFICIAL INTELLIGENCE (AI), AND PASSIVE PROBLEM-SOLVERS

Given this definition, an **Artificial Intelligence** (or AI) is a **purposefully-engineered** system that **solves problems**. For our purposes, we will make a distinction between (purely cognitive) **passive problem-solvers** and (agentic) **active problem-solvers**. A *passive problem-solver* takes as input some kind of *problem statement*, and generates a *problem solution* as output. As a very simple example, the problem statement might be "what is 1 + 1?", and the problem solution might be "2".

In contrast, an *active problem-solver* (or **agent**) takes as input a *continuous stream of percepts*, and performs a *continuous stream of actions* as output, all in pursuance of its (fixed) *final goal*[5].

2.2.8 PROBLEM STATEMENTS

For our purposes, the ideal **problem statement** has three parts (expressed as a *triple* $\langle P, Q, R \rangle$):

- P: a property over all possible things T
- Q: an (optional) ordering over all possible things T having property P
- R: a constraint on the use of physical resources (always including a finite time limit).

Simply stated:

- P tells us which things T are **valid solutions** (there may be 0, 1, or many) and which are not
- if more that one valid solution exists, Q tells us which valid solutions are **better** than others[6]
- R tells us how much time, energy, compute, money etc we can use in the search for a solution[7].

A triple $\langle P, Q, R \rangle$ should be interpreted by a passive problem-solver as follows: "**using no more than physical resources R, strive to find some Q-maximal thing x having property P**"; e.g.[8]:

- using no more than 60 s of time or 1 MJ of energy, strive to find some thing x such that x is a natural number and $x = 1 + 1$ — in this case, there is only one solution (i.e. $x = 2$)
- using no more than 60 s of time or 1 MJ of energy, strive to find some Q-maximal thing x such that x is an integer and $x^2 = 4$, where Q favours positive solutions — in this case, there are two possible solutions ($x = -2$ and $x = 2$); however, because the ordering Q favours positive solutions, $x = 2$ is considered to be a "better" (higher quality) solution than $x = -2$.

In principle, problems $\langle P, Q, R \rangle$ may be **arbitrarily complex**, i.e. as complex as the real world.

[5] active problem-solvers are described more fully in Section 2.2.29
[6] if Q is not specified, then any valid solution is considered to be as good as any other
[7] in particular, in many real-world contexts, only *timely* solutions have any significant utility [see e.g. Newell (1990)]
[8] at this point in the narrative we are more focused on *intuitive concepts* than *formal definitions*, and so, for simplicity, problems $\langle P, Q, R \rangle$ are described in English; in later sections $\langle P, Q, R \rangle$ may be expressed in a more formal notation

2.2.9 PROBLEM SOLUTIONS

Given an arbitrary problem $\langle P, Q, R \rangle$, a passive problem-solver:

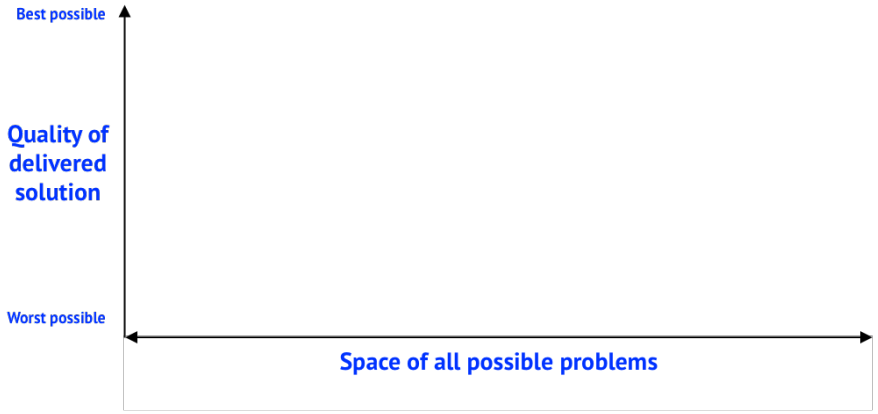
- will **strive to find the best valid solution that it can**, with no guarantees beyond that
- may deliver any of the following[9]:
  - the best possible (i.e. maximum) solution, according to the specified (total) ordering
  - a best (i.e. maximal) solution, according to the specified (partial) ordering
  - a satisfactorily good solution, according to the specified ordering, albeit not the best
  - a disappointingly poor solution, according to the specified ordering, albeit not the worst
  - a worst possible (i.e. minimal) solution, according to the specified (partial) ordering
  - the worst possible (i.e. minimum) solution, according to the specified (total) ordering
  - no solution at all (e.g. if the problem has no solution, or if the AI runs out of resources).

When dealing with **arbitrary problems** of the form $\langle P, Q, R \rangle$, this is the best we can hope for[10].

Some passive problem-solvers will deliver **better** (i.e. **higher-quality**) solutions than others.

2.2.10 SOLUTION QUALITY

If we set the finite time limit within which problem solutions must be found to some fixed value, but otherwise allow a problem-solver unbounded physical resources R, then we're left with two dimensions, which we can visualise as a simple 2D graph showing (1) the specific **problem** defined by P and Q (on the x-axis), and (2) the **quality** of the delivered solution according to ordering Q (on the y-axis). We can then plot different passive problem-solvers on the same 2D graph, for comparison[11]:



(For the avoidance of doubt, the following graphs (depicted in Sections 2.2.11 to 2.2.17) pertain specifically to *passive problem-solvers* (as defined in Section 2.2.7), assuming a *fixed time limit*.)

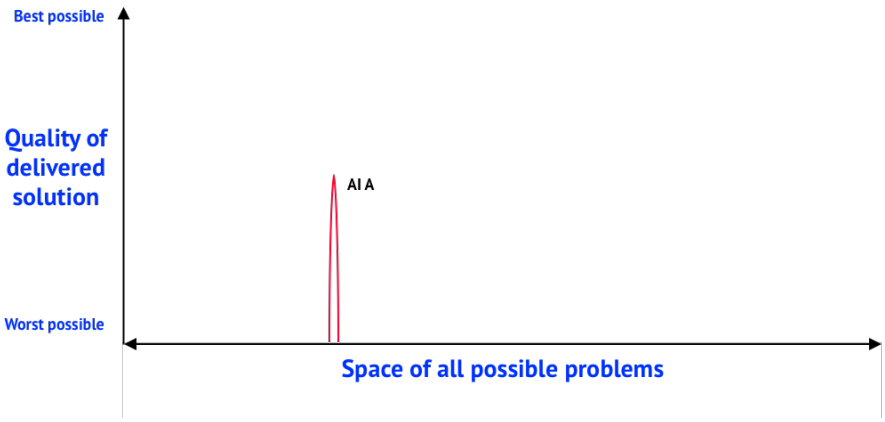
[9] for our present purposes, we make the simplifying assumption that any passive problem-solver is **deterministic** (given the same $\langle P, Q, R \rangle$, it always delivers the same result); in reality, many passive problem-solvers will be **nondeterministic**

[10] note that if we were to give an arbitrary problem to a diligent human, the possible outcomes would be exactly the same!

[11] such diagrams are intended for illustrative purposes only, and should not be interpreted as any kind of formal definition

2.2.11 NARROW AI

A **narrow AI** delivers valid solutions across a narrow range of problems:

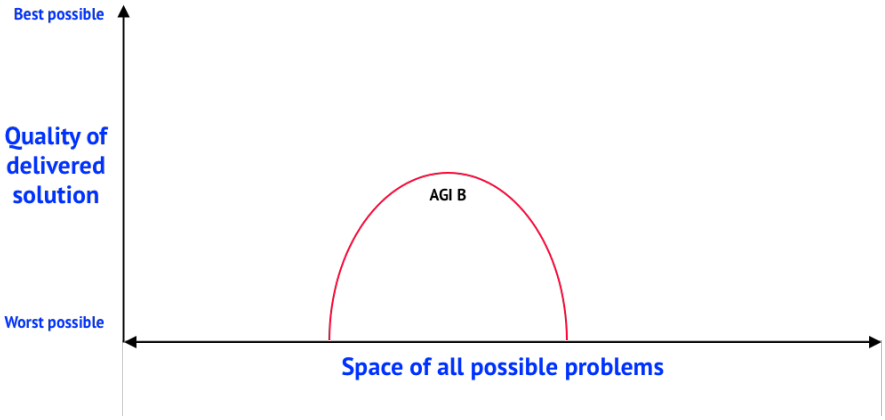


Here, AI A is a narrow AI (according to our definition).

Note that we do not require a narrow AI's delivered solutions to exceed any minimum level of quality, beyond being valid solutions according to the specified problem statement. Thus this is a very broad definition, and some very weak problem-solvers will nevertheless qualify as narrow AI.

2.2.12 ARTIFICIAL GENERAL INTELLIGENCE (AGI)

A **general AI** (GAI, usually styled **AGI**), delivers valid solutions across a wide range of problems:



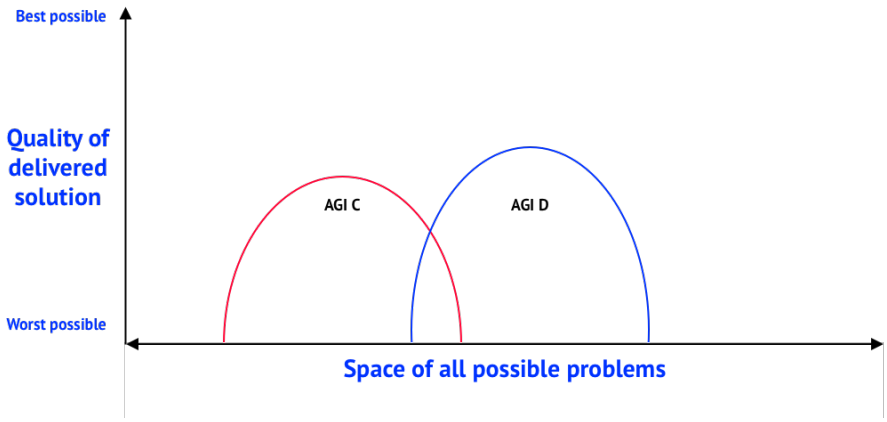
Here, AGI B is an AGI (according to our definition).

Again, we do not require an AGI's delivered solutions to exceed any minimum level of quality.

We also do not attempt to define the threshold between narrow AI and AGI, other than to say that a narrow AI delivers valid solutions across a **single problem domain**, whereas an AGI delivers valid solutions across **multiple problem domains**. It will usually be intuitively clear what a "problem domain" is. Nevertheless, the distinction between narrow AI and AGI remains somewhat subjective.

2.2.13 AGI COMPETENCE

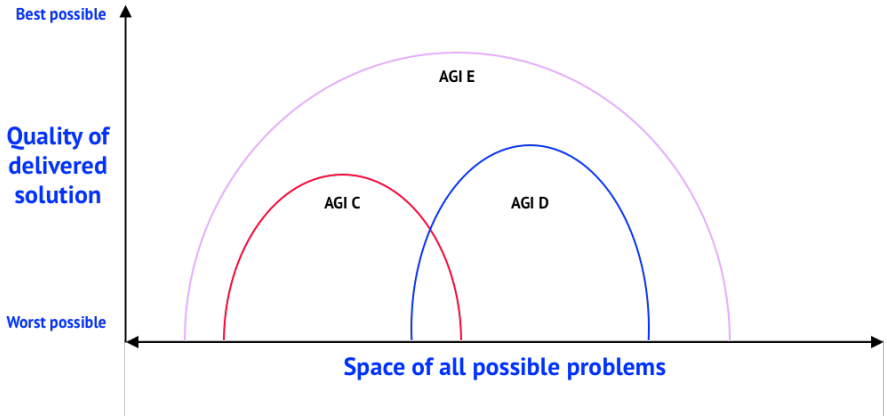
Some AGIs might be particularly adept in some problem areas, but not in others:



Here, AGI C and AGI D are good at different things.

2.2.14 THE ALPHA AGI

It's possible for one AGI — the **alpha** — to outperform all its peers across all possible problems:

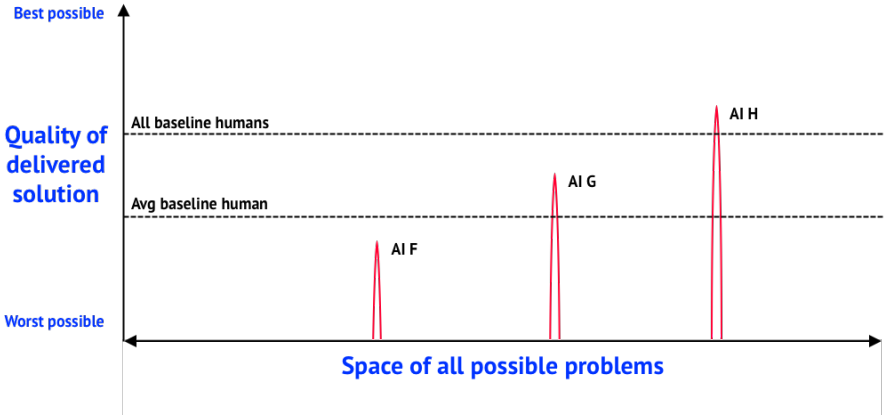


Here, AGI E outperforms both AGI C and AGI D at all things.

The concept of the alpha AGI will become extremely important later on.

2.2.15 SUPERINTELLIGENT NARROW AI

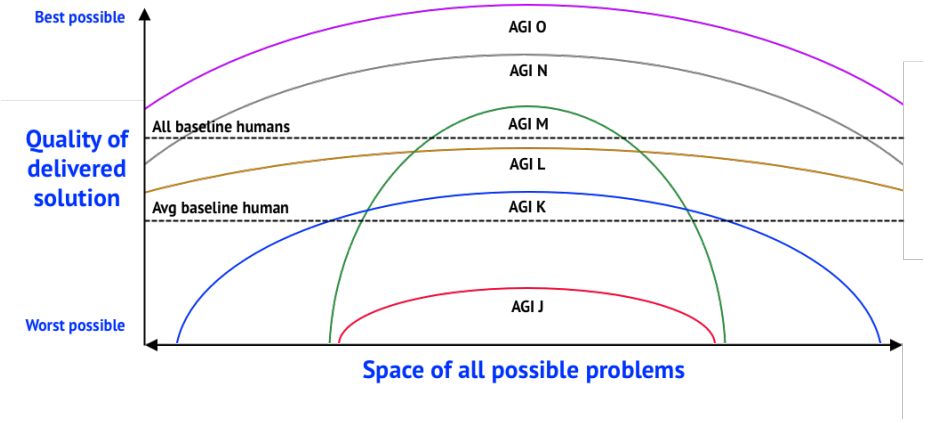
Things get interesting when we compare **AI** and **baseline human** (BH) performance[12]:



Here, the average BH outperforms narrow AI F, narrow AI G outperforms the average BH, and narrow AI H outperforms all BHs (i.e. the entire population of BHs). In other words, narrow AI H (e.g. a modern pocket calculator) is **superintelligent** across the narrow range of problems in question.

2.2.16 SUPERINTELLIGENT AGI

We can also compare **AGI performance** against **BH performance**:

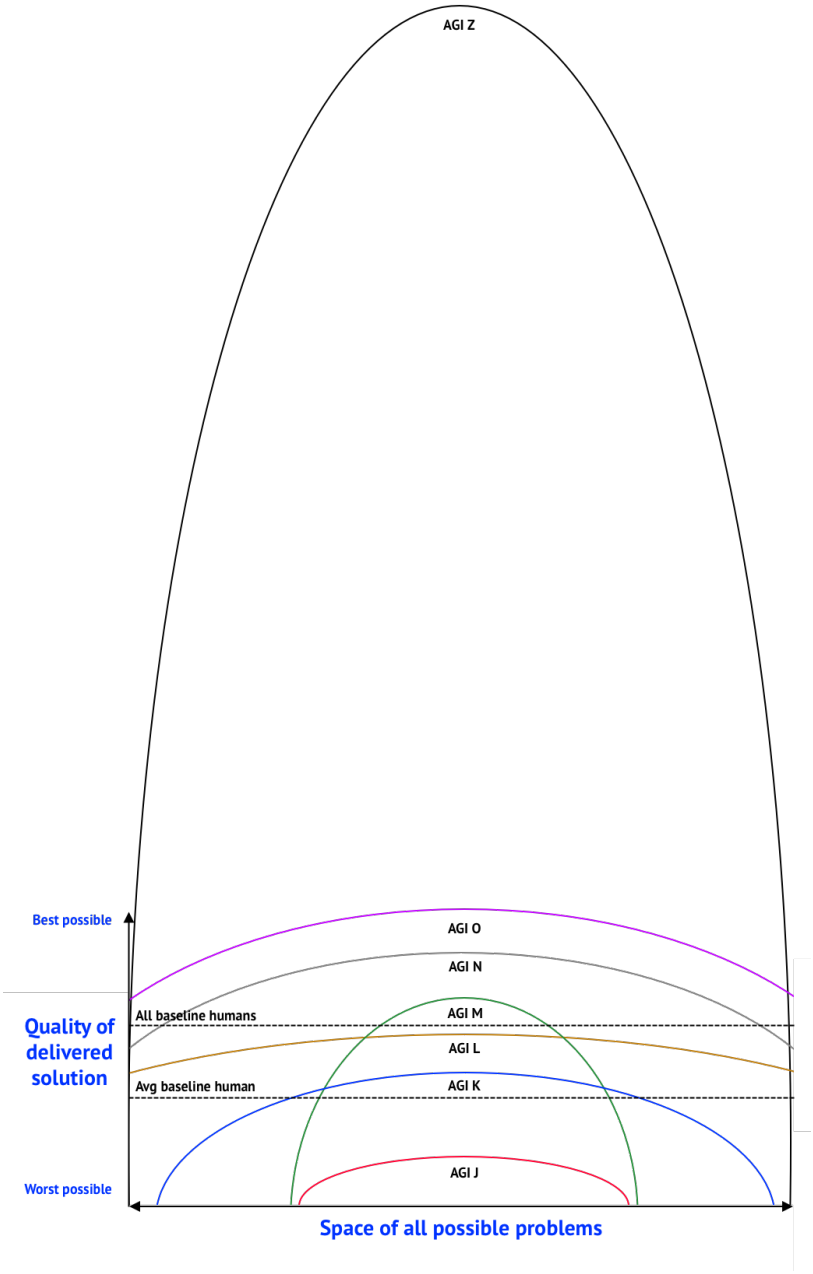


There are six primary possibilities. The average BH outperforms AGI J at all things, AGI K outperforms the average BH at some things[13], AGI L outperforms the average BH at all things[14], AGI M outperforms all BHs at some things[13] and the average BH at as least as many things, AGI N outperforms all BHs at some things[13] and the average BH at all things, and AGI O outperforms all BHs at all things. In other words, AGI O is **superintelligent** across all possible problems[15].

[12] by "baseline human" we mean a human unassisted by any other system that would itself qualify as an AI as defined above
[13] note: *across multiple problem domains* — exceeding the threshold for a single problem domain is not sufficient!
[14] broadly speaking, AGI L seems to be what most AI researchers mean by **human-level AGI** [e.g. Morris et al. (2023)]
[15] note that a sufficiently performant AGI M or AGI N might be judged to be **near-superintelligent**

2.2.17 MAXIMALLY-SUPERINTELLIGENT AGI

Of course, there's no need to stop at superintelligence. It might be possible to go way beyond AGI O:



Intuitively, a **maximally-superintelligent** AGI is the most intelligent superintelligent AGI that it's possible to build. Merely outperforming all baseline humans at everything is a lower bound.

2.2.18 EXAMPLE — GPS

The **General Problem-Solving Program** (GPS) [Newell, Shaw, and Simon (1958); Ernst and Newell (1969)] was an early attempt at a general problem-solving system. Once a problem description had been input to the system[16], GPS would attempt to find a path to a solution via *means-ends analysis*[17].

Although only ever intended as an exploratory research project, and heavily constrained by the limits of 1960s computer systems, GPS could potentially be applied to a wide range of problems.

In the course of its roughly 10-year existence, GPS was applied to problems such as the following:

- **Missionaries and Cannibals** — Three missionaries and three cannibals wish to cross a river, but their boat only holds two people; how can all six get across with no-one being eaten?
- **Symbolic integration** — Integrate $\int te^{t^2}\,dt$.
- **Tower of Hanoi** — Discover a sequence of moves that will transfer all the disks (all of different sizes) from the first peg to the third peg, such that no disk is ever on top of a smaller disk.
- **First-order logic theorem-proving** — Use the resolution rule [Robinson (1965)] to prove: $\exists u \exists y \forall z\,((P(u,y) \Rightarrow (P(y,z) \wedge P(z,z))) \wedge ((P(u,y) \wedge Q(u,y)) \Rightarrow (Q(u,z) \wedge Q(z,z))))$.
- **Father and Sons** — A father (weighing 200 pounds) and his sons (each weighing 100 pounds) wish to cross a river in a boat whose capacity is 200 pounds; how can they all get across?
- **Monkey** — A room contains a monkey, a box, and some bananas hanging from the ceiling, but they are too high for the monkey to reach; how can the monkey get the bananas?
- **Three coins** — Make three coins (initially T, H, T) all show the same, in just three moves.
- **Parsing sentences** — Correctly parse the sentence "Free variables cause confusion".
- **Bridges of Königsberg** — The river Pregel runs through the German town of Königsberg[18]; in the river are two islands connected with the mainland and with each other via seven bridges; is it possible to cross each of the seven bridges exactly once and return to the same point?[19]
- **Water jug** — Given a water tap, a drain, a five-gallon jug, and an eight-gallon jug (and no other water-measuring devices), how can exactly two gallons of water be put into the five-gallon jug?
- **Letter series completion** — Complete the series "B C B D B E __ __".

GPS was able to successfully solve all of the above problems, except for the Bridges of Königsberg.

Is GPS intelligent? From the perspective of an external observer, GPS takes as input a *problem statement*, and generates a *problem solution* as output. According to our earlier definitions, therefore, **GPS is a passive problem-solver demonstrating extrinsic intelligence**.

Is GPS an AGI? GPS is not limited to a single problem domain, so **GPS is an AGI**.

Is it superintelligent? No — we would judge GPS to be an **AGI K**, far short of AGI O.

[16] in terms of the *objects* pertinent to the problem in question and the *operators* that may be applied to them

[17] effectively working backwards from the desired final state, via the available operators, to the initial state

[18] now Kaliningrad in Russia

[19] in 1736, the mathematician Leonhard Euler proved this to be impossible

2.2.19 EXAMPLE — DALL·E

DALL·E [Ramesh et al. (2021)] generates an *image* in response to a *text prompt*[20].

For example, the following images were generated by DALL·E:

Is DALL·E intelligent? From the perspective of an external observer, DALL·E takes as input a *problem statement*, and generates a *problem solution* as output. According to our earlier definitions, therefore, **DALL·E is a passive problem-solver demonstrating extrinsic intelligence**.

Is DALL·E an AGI? DALL·E is limited to a single problem domain, so **DALL·E is not an AGI**.

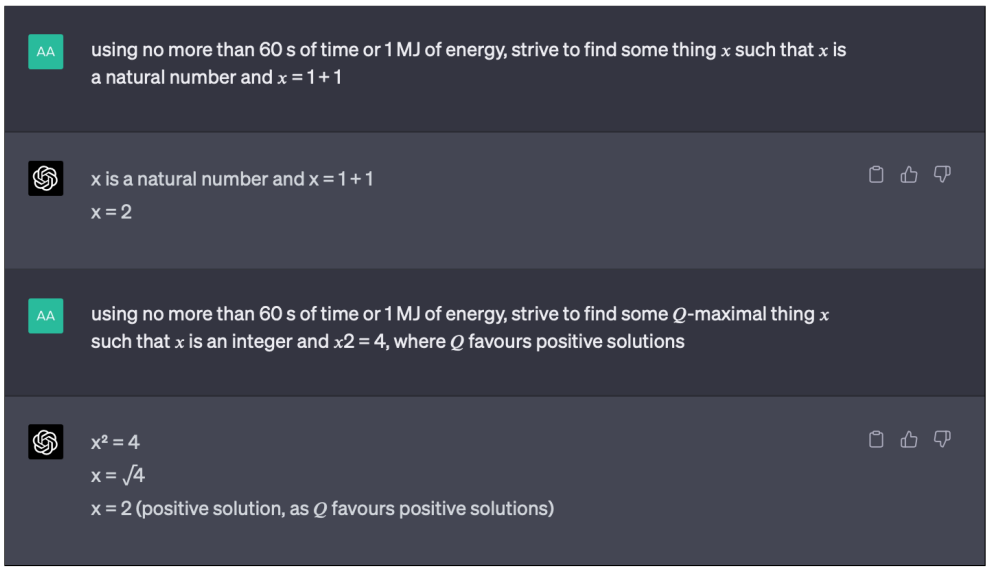
Is it superintelligent? No — we would judge DALL·E to be an **AI G**, far short of AI H.

[20] *Midjourney* [Oppenlaender (2022)] is a similar text-to-image system

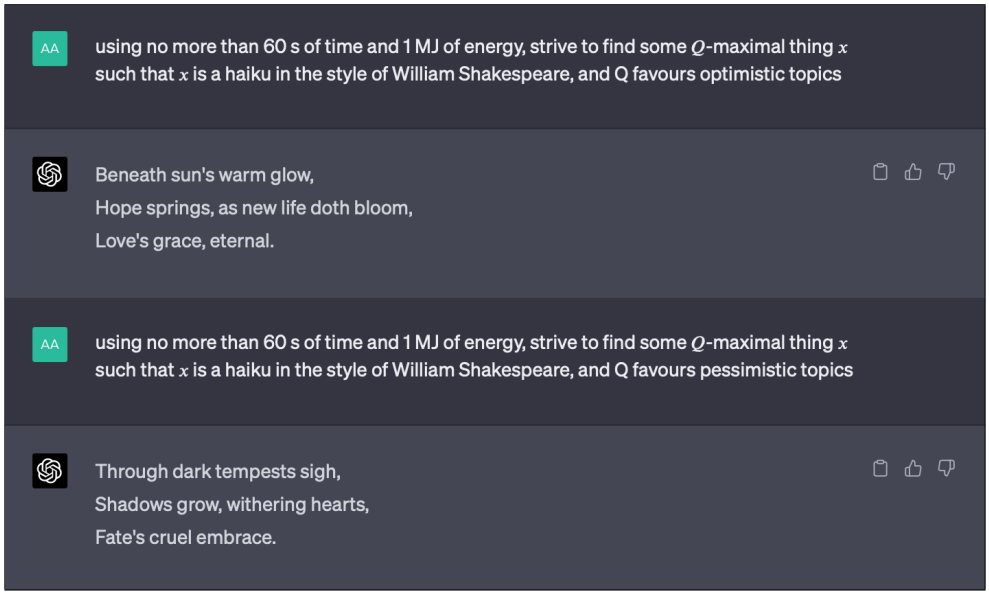
2.2.20 EXAMPLE — CHATGPT

ChatGPT[21, 22] [Brown et al. (2020); Liu et al. (2023)] generates *text* in response to a *text prompt*.

Example — arithmetic:

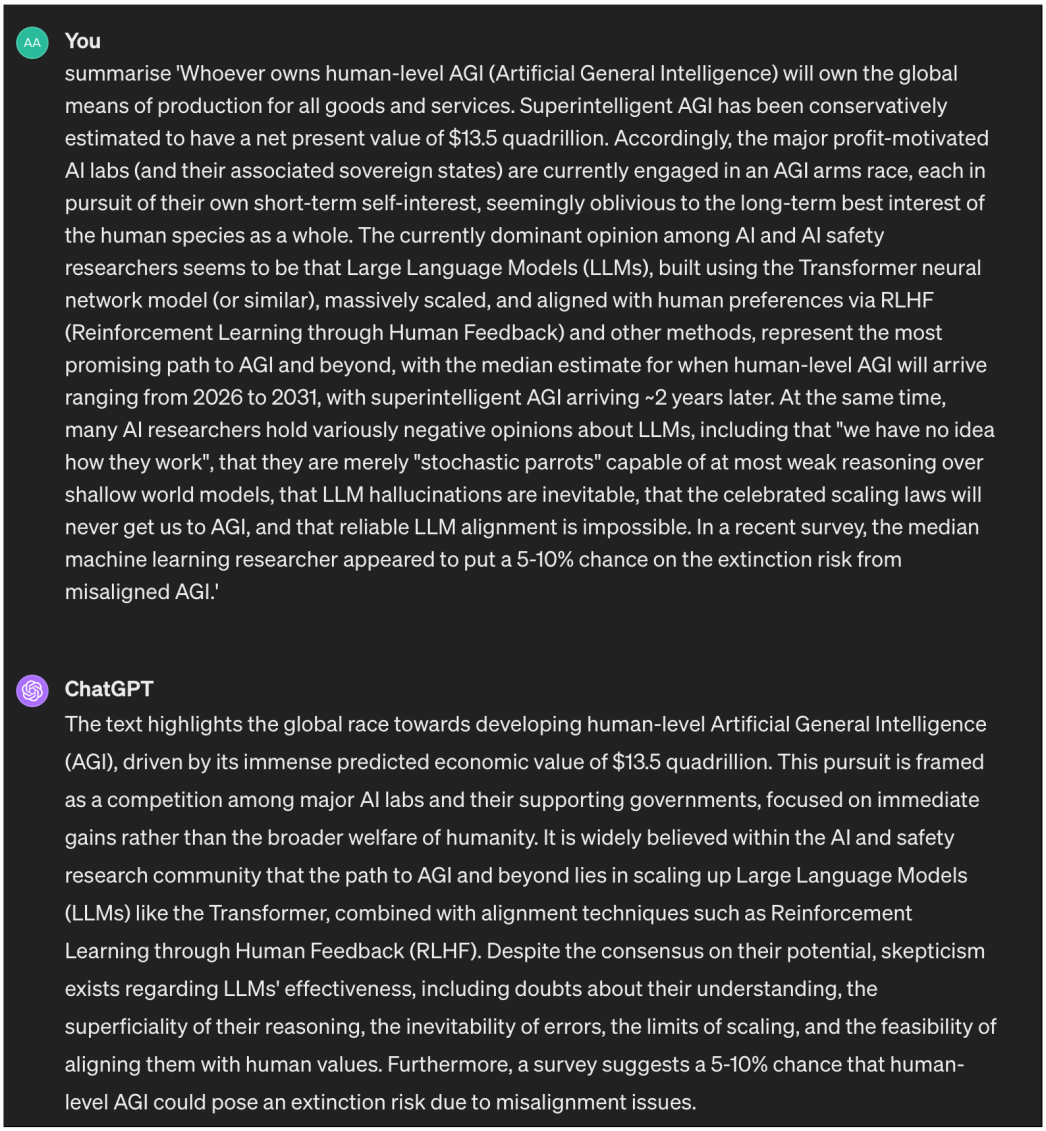


Example — haikus[23]:



[21] specifically, the mid-2023 version based on the GPT-4 **Large Language Model** (LLM), without any external embellishments
[22] note that the Bard chatbot (powered by the Gemini 1.0 LLM released by Google DeepMind in December 2023) is broadly comparable, as is the Claude 3 LLM released by Anthropic in March 2024 [see Romero (2024) for an informal analysis]
[23] note that the pessimistic haiku only works if cruel is pronounced "cru-el"

Example — summarisation:



Is ChatGPT intelligent? From the perspective of an external observer, ChatGPT takes as input a *problem statement*, and generates a *problem solution* as output[24]. According to our earlier definitions, therefore, **ChatGPT is a passive problem-solver demonstrating extrinsic intelligence**.

Is ChatGPT an AGI? ChatGPT is not limited to a single domain, so **ChatGPT is an AGI**.

Is it superintelligent? No — we would judge ChatGPT to be an **AGI K**[25], far short of AGI O.

[24] although ChatGPT generates surprisingly good solutions in many cases, its responses are often riddled with factual inaccuracies; for our purposes, we consider factually inaccurate problem solutions to be invalid, and disregard them

[25] i.e. broadly the same as GPS in terms of solution quality, only much easier to use, and serving a wider range of problems; note that *four* AI systems, 60 years apart, all stalling at AGI K, suggests that AGI K is a difficult barrier to overcome!

2.2.21 INTRINSIC INTELLIGENCE

Following our earlier definitions, we have determined that both DALL·E and ChatGPT are passive problem-solvers exhibiting extrinsic intelligence, and we have also determined that ChatGPT is an AGI, albeit neither a human-level nor a superintelligent AGI. But is extrinsic ("black box") intelligence a sufficient measure of *genuine* intelligence? Or is whatever happens "inside the box" also relevant?

Imagine if we opened up DALL·E or ChatGPT and found nothing inside but a massive lookup table going from problem statements to problem solutions. If this were the case, then the system's **externally observable** behaviour would be identical to what we have already judged to be **extrinsically intelligent**, yet we would not judge such a system to be **intrinsically intelligent**[26]. Therefore:

- a system's **internal behaviour** is highly relevant to whether or not it is intrinsically intelligent
- we cannot reliably infer a system's internal behaviour just by observing its **external behaviour**[27]
- if we want to know what's going on inside the box, we have to actually look inside the box!

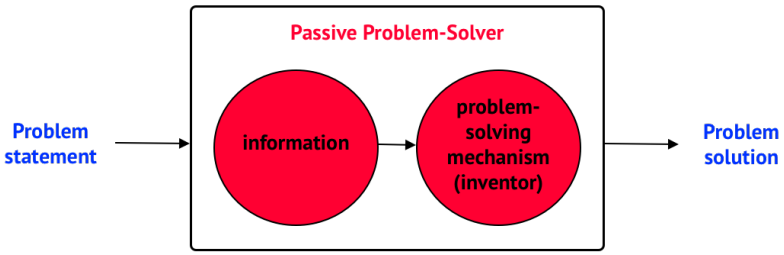
2.2.22 THE ROLE OF INFORMATION IN PROBLEM-SOLVING

If we look inside a passive problem-solver, we will see some kind of **problem-solving mechanism** (e.g. software). But **problem-solving is not just about algorithms** [Wirth (1976); Kowalski (1979)].

If there's one key takeaway to be gleaned from the classic problem-solving literature [e.g. Pólya (1945); Pólya (1954a); Pólya (1954b); Newell, Shaw, and Simon (1958); Pólya (1962a); Pólya (1962b); Ernst and Newell (1969); Newell and Simon (1972); Lakatos (1976); Newell (1990); Tsang (1993); Dechter (2003)], it's that **the key to effective problem-solving is the effective use of information**.

Thus, when we look (closely) inside a passive problem-solver, we will necessarily see *two* things:

1. the underlying **problem-solving mechanism** per se (which we shall call the **inventor**[28])
2. the **information** (*data*, *knowledge*, *belief*[29]) that **drives** the inventor towards solutions:



For example, when searching the space of all possible values of x trying to find a solution to the equation $x = 1 + 1$, the inventor may be guided by its knowledge of the number 1 and the operator $+$.

[26] the famous *Chinese Room Argument* [Searle (1980); Cole (2023)] makes essentially the same observation

[27] therefore, just because DALL·E and ChatGPT possess *extrinsic intelligence*, it's not necessarily the case that they possess significant *intrinsic intelligence* — any such assertion based solely on externally observable behaviour (such as that described above) is merely *one* possible explanation of that externally observable behaviour (i.e. an abductive hypothesis)

[28] from the Latin *invenire*: to find, discover, invent, devise

[29] for our purposes, we shall use these terms interchangeably — philosophers should have a nap and a banana at this point! :-)

The **information** that drives the inventor towards solutions must be:

- **acquired** (externally) somehow[30]
- **represented** (internally) somehow[31]
- **processed** (by the inventor) somehow[32].

2.2.23 DON'T PANIC!

> It is said that despite its many glaring (and occasionally fatal) inaccuracies, the *Hitchhiker's Guide to the Galaxy* itself has outsold the *Encyclopedia Galactica* because it is slightly cheaper, and because it has the words 'DON'T PANIC' in large, friendly letters on the cover. Adams (1978)

The next few Sections, from 2.2.24 to 2.2.26, require a little mental gymnastics on the part of the reader. Considering that Part 1 of the BigMother series of papers (i.e. this Part) aims to be as **accessible** to non-technical readers as possible, this is the conceptually most difficult part of the paper to read, for which we apologise, but there doesn't seem to be any way around it. Once one of our informal reviewers, Professor Bob Kowalski, had pointed out a glaring inconsistency in a later part of the paper, we had no choice but to resolve it, and Sections 2.2.24 to 2.2.26 are the result. The paper, we believe, has been significantly strengthened overall (thanks Bob!), with the minor downside that some readers may struggle to read these particular sections as much as we struggled to write them!

2.2.24 MAXIMALLY-INTRINSIC PASSIVE PROBLEM-SOLVERS

Earlier, in Section 2.2.21, we determined that both DALL·E and ChatGPT are passive problem-solvers exhibiting extrinsic intelligence. We also imagined the possibility of opening up either DALL·E or ChatGPT and finding nothing inside but a massive lookup table going from problem statements to problem solutions. In such a configuration, the **inventor** (the underlying problem-solving mechanism per se) is merely an extremely simple algorithm that uses the problem statement as a *key* to index into the massive lookup table, delivering the contents of the table row so accessed as the problem solution, and the massive lookup table is the **information** that drives that inventor towards solutions.

As already alluded, even though such a system exhibits **extrinsic intelligence**, it lacks **intrinsic intelligence**. The *real* intelligence resides in the **human AI designer**[33] that created the lookup table!

In constructing the lookup table, the human AI designer would need to:

1. consider every possible problem statement (i.e. every possible problem instance)
2. attempt to solve each problem instance themselves (e.g. by hand-executing a suitable algorithm)
3. add the results to the lookup table.

[30] passive problem-solvers don't do any *information acquisition* themselves (it's all done for them by some other party)
[31] for a wannabe superintelligent AGI, the underlying **knowledge representation language** (its **language of thought** [Kowalski (2011a); Rescorla (2019)]) must be as *general* as possible — otherwise there may be concepts that a human being can express (and thereby reason about) but that the AGI cannot, immediately rendering superintelligence impossible
[32] e.g. via an arbitrarily complex combination of strong probabilistic *induction*, *deduction*, and *abduction* (2.2.27 and 2.2.28)
[33] for the time being, we shall mostly restrict our discussion to the case where a *human being* is designing an *AI system*

In other words, the AI designer is operating as a "higher-level" (or **meta-level**) problem-solver[34], and the passive problem-solver being designed is the "lower-level" (or **object-level**) problem-solver[35]:

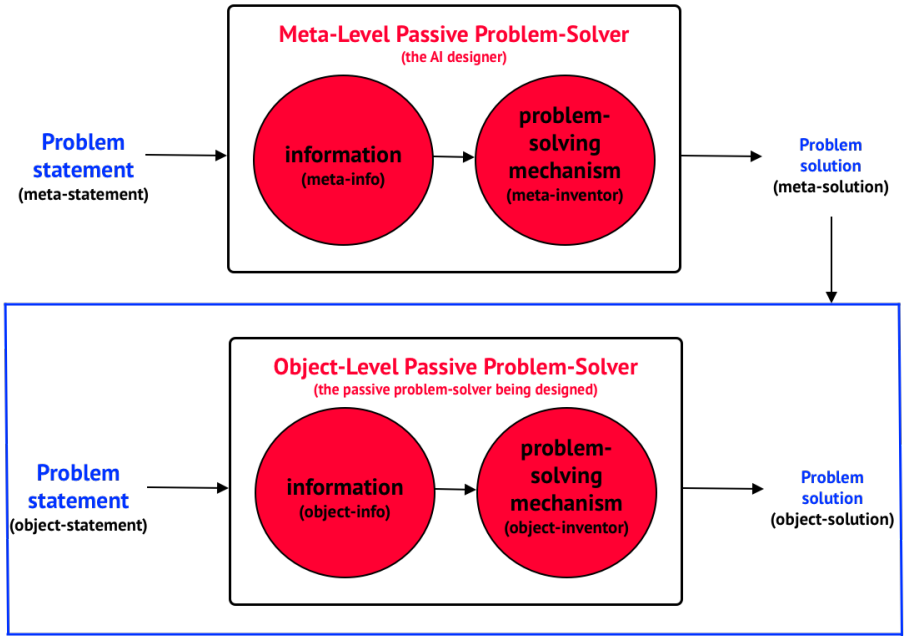


Figure 1: A meta-level passive problem-solver designing an object-level passive problem-solver

The components of this *dual-problem-solver* system are:

meta-statement	the meta-level problem statement (possibly in some natural language)
meta-info	other information that the meta-level problem-solver has in its possession
meta-inventor	the meta-level problem-solving mechanism (for example, a human mind)
meta-solution	the meta-level problem solution (i.e. the solution to the meta-statement)
object-statement	the object-level problem statement (expressed in a specific language or format)
object-info	other information that the object-level problem-solver has in its possession
object-inventor	the object-level problem-solving mechanism (e.g. some AI algorithm)
object-solution	the object-level problem solution (i.e. the solution to the object-statement)

We define **pre-calculated derived information** as any information derived from any object-statement that is either (a) contained in object-info, or (b) encoded into object-inventor.

We will say that an object-level passive problem-solver is **maximally-intrinsic**[36] if object-inventor is capable of calculating any object-solution that may be calculated *using* pre-calculated derived information (e.g. a lookup table) *without using* any pre-calculated derived information.

[34] it is sometimes useful to view *humans* as problem-solvers (either active or passive)
[35] in reality, there may be more than two levels (such as a top-level AI designer who designs a mid-level AI that generates information for the lowest-level AI), with each higher level acting as the meta-level problem-solver for the level below
[36] short for **maximally-intrinsically-intelligent**

The net effect of ensuring that an object-level passive problem-solver is maximally-intrinsic is to move **all information** and **all calculation** on the computational path from object-statement to object-solution from the meta-level passive problem-solver into the object-level passive problem-solver[37].

Hopefully an example will help to clarify. In the "massive lookup table" example:

- within the meta-level problem-solver (the AI designer, modelled as a passive problem-solver):
    1. meta-info contains information pertaining to mathematics, computer science, AI design, etc, such as how to construct a passive problem-solver structured as a lookup table
    2. meta-inventor initialises the lookup table such that every row contains "no solution found"
    3. from meta-statement conjoined with meta-info, meta-inventor derives an algorithm $\mathcal{A}$, which strives to find a valid object-solution for any given instance of object-statement
    4. meta-inventor simulates algorithm $\mathcal{A}$ (thereby yielding a valid object-solution, assuming that one can be found for the problem instance in question) for every object-statement (i.e. problem instance) that it is able to derive from meta-statement conjoined with meta-info, inserting any object-solution so found into the corresponding row of the lookup table — note that every object-solution so found has been calculated (not by the object-level passive problem-solver, but by the meta-level problem-solver) from the corresponding object-statement, and therefore qualifies as **pre-calculated derived information**
    5. from meta-statement conjoined with meta-info, meta-inventor derives object-inventor (a simple algorithm to index into a lookup table, such as the one just constructed)
    6. the meta-solution is (an implementable design for) the object-level problem-solver (comprising (a) the completed lookup table as object-info, and (b) object-inventor)
- within the object-level problem-solver (an operational implementation of the meta-solution):
    1. object-info contains the lookup table (i.e. a collection of pre-calculated object-solutions)
    2. object-inventor uses object-statement to index into the lookup table, yielding (**without any further calculation**) either "no solution found" or the corresponding object-solution
- thus:
    (a) the lookup table contained in object-info contains pre-calculated derived information
    (b) object-inventor calculates object-solutions using that pre-calculated derived information
    (c) ... but is unable to do so *without* using any pre-calculated derived information
    (d) accordingly, the object-level passive problem-solver in question is *not* maximally intrinsic.

Note that it would also be possible for the meta-level problem-solver to achieve equivalent effect not by generating a lookup table for explicit incorporation into object-info but by instead structuring the object-inventor (algorithm) as a big "switch" statement, effectively encoding the lookup table into the object-inventor. This would of course also qualify as pre-calculated derived information.

[37] in other words, a **maximally-intrinsically-intelligent** (i.e. maximally-intrinsic) problem-solver **does all its own thinking**; accordingly — because some of its thinking is being done for it (by a meta-level problem-solver) — a *non*-maximally-intrinsic passive problem-solver can't possibly fully understand its own reasoning, whereas — because it *does* do all its own thinking — a maximally-intrinsic passive problem-solver must *necessarily* fully understand its own reasoning

2.2.25 MAXIMALLY-INTRINSIC SUPERINTELLIGENT AGI

A **maximally-intrinsic superintelligent passive problem-solver** (AGI O/Z) must necessarily possess:

1. an inventor that is *better at solving any given problem than any human*
2. *a deep understanding of all human knowledge* (grounded by experience)[38]
3. sufficient physical resources (e.g. compute) to be able to deliver *timely* problem solutions.

We shall call these qualities *super-inventive*, *super-knowledgeable*, and *super-resourced*. Thus:

superintelligent = super-inventive + super-knowledgable + super-resourced[39].

2.2.26 THE PATH TO SUPERINTELLIGENT AGI

In general, **intelligence has three scalable dimensions**: *inventiveness + knowledge + resources*[40].

Accordingly, if a human AI designer (meta-level passive problem-solver) $\mathcal{H}$ designs a *non-maximally-intrinsic* object-level passive problem-solver $\mathcal{M}-$ then that means that some sub-components of $\mathcal{M}-$ (and thereby the quality of object-solutions delivered by $\mathcal{M}-$) are partly reliant on pre-calculated derived information $\mathcal{DI}$ generated by the *inventiveness + knowledge + resources* combination of $\mathcal{H}$'s meta-inventor + $\mathcal{H}$'s meta-information + "one human brain's worth of compute".

Conversely, if $\mathcal{H}$ instead designs a *maximally-intrinsic* variant of $\mathcal{M}-$ called $\mathcal{M}+$ then $\mathcal{DI}$ (on which the quality of object-solutions delivered by $\mathcal{M}+$ depends) may be generated by $\mathcal{M}+$'s object-inventor + $\mathcal{M}+$'s object-information + whatever physical resources $\mathcal{M}+$ has available to it.

Comparing the three passive problem-solvers in question ($\mathcal{H}$, $\mathcal{M}-$, and $\mathcal{M}+$):

- $\mathcal{H}$'s meta-inventor, $\mathcal{M}-$'s object-inventor, and $\mathcal{M}+$'s object-inventor are (assumed to be) fixed
- $\mathcal{H}$'s meta-information scales relatively slowly with time (reading literature, taking courses, etc)
- both $\mathcal{M}-$'s and $\mathcal{M}+$'s object-information can potentially scale much more rapidly
- $\mathcal{H}$'s physical (compute) resources are fixed (at ~3 pounds of grey matter and ~20 W of power)
- $\mathcal{M}-$'s and $\mathcal{M}+$'s physical resources can potentially be scaled by several orders of magnitude[41].

Thus the potential scope for improving the performance (object-solution quality) of either $\mathcal{M}-$ or $\mathcal{M}+$ by scaling $\mathcal{H}$'s meta-information or physical resources is limited by the rate at which $\mathcal{H}$ can gain new knowledge. There is greater scope for improving the performance of $\mathcal{M}-$ by scaling $\mathcal{M}-$'s object-information or physical resources, but, even if $\mathcal{M}-$ were near-superintelligent, this would *not* improve the quality of the pre-calculated derived information $\mathcal{DI}$ on which the quality of object-solutions delivered by $\mathcal{M}-$ is partly reliant. Thus the **potential scope for improving performance via scaling** is greatest for near-superintelligent $\mathcal{M}+$, because in this case the quality of $\mathcal{DI}$ might also be improved. Therefore, given equivalent near-superintelligent $\mathcal{M}-$ and $\mathcal{M}+$, $\mathcal{M}+$ has the greatest chance of being pushed across the line from **sub-superintelligence** to **superintelligence** via scaling.

[38] a **deep understanding** of any subject provides much more **problem-solving information** than a **shallow understanding**
[39] *maximally-superintelligent = maximally-super-inventive + maximally-super-knowledgeable + maximally-super-resourced*
[40] note that the actual relationship between these qualities is more likely to be multiplicative than additive!
[41] a modern supercomputer weighs ~600,000 pounds and consumes 20-60 MW of power [Dongarra and Geist (2022)]

2.2.27 INDUCTION, DEDUCTION, ABDUCTION, AND WEAK VS STRONG REASONING

It seems to be the case that, at the highest levels of abstraction (at the level of *mind* rather than *brain*), **robust critical thought and problem-solving within humans** involves three modes of reasoning[42]:

- **induction**[43, 44] — the discovery of *patterns* — Example 1: after seeing a number of examples of swans (all of which are white), you formulate the general concept of "swan" (and, based on your experience, all swans are white); Example 2: after seeing a number of examples of cats (all of which have tails), you formulate the general concept of "cat" (and, based on your experience, all cats have tails); Example 3: after learning of the existence of black swans (which are not white) and Manx cats (which don't have tails), you realise that not all swans are white, and not all cats have tails, and you formulate the abstract concept of "exception"; Example 4: after seeing a number of dead swans, cats, and men, you formulate the abstract concept of "mortal"
- **deduction**[45] — the derivation of *necessary conclusions* — Example: Socrates is a man, and all men are mortal, therefore [via the *modus ponens* rule of inference] Socrates is mortal
- **abduction**[46] — the formulation of *possible explanations* — Example: Socrates is mortal, therefore Socrates could be a swan, Socrates could be a cat, and Socrates could be a man[47].

Thus induction, deduction, and abduction (IDA) may be viewed as **cognitive primitives** on top of which higher-level reasoning such as **generic problem-solving** might be constructed. Specific formulations of IDA may be either *strong* (well-founded) or *weak* (prone to some kind of error)[48]:

- *weak* induction, e.g. either seeing patterns that aren't in the data, or not seeing patterns that are
- *weak* deduction, e.g. concluding something that isn't necessarily a consequence (*non sequitur*)
- *weak* abduction, e.g. elevating a plausible hypothesis to an unqualified belief, without evidence.

We conjecture the following[49]:

1. these 3 modes of reasoning, suitably integrated, are *necessary* for ≥ human-level AGI (AGI L)
2. these 3 modes of reasoning, suitably integrated, are *sufficient* for ≥ human-level AGI (AGI L)
3. any AGI that is able to perform *strong* induction, deduction, and abduction will have an advantage over an AGI that is limited to relatively *weak* induction, deduction, and/or abduction (as the weaker forms will introduce errors of reasoning, leading to inferior problem solutions)
4. superintelligent AGI (AGI O/Z) will require *strong* induction, deduction, and abduction.

[42] NB it is common to see conflicting definitions of *induction* and *abduction* in the literature [Flach and Hadjiantonis (2000)]
[43] Holland et al. (1986); Mortimer (1988); Flach and Hadjiantonis (2000); Johnson (2016); Henderson (2020); Bartha (2022)
[44] note that *induction* and *analogy* are closely related, a subject to which we shall return in Section **??**
[45] Shoenfield (1967); Barwise (1977); Mendelson (1987); Halbach (2010)
[46] Flach and Hadjiantonis (2000); Walton (2014); Douven (2021); Douven (2022)
[47] these are *plausible hypotheses* only; elevation to *unqualified belief* (e.g. Socrates *is* a cat) requires additional evidence
[48] humans, of course, are prone to a wide range of errors of reasoning, as we shall discuss in Section **??**
[49] it will likely take many decades of research for these conjectures to be tested, but nevertheless these are our suspicions

2.2.28 UNCERTAINTY

As we shall see in Section **??**, any belief pertaining to the (assumed) physical universe (including any belief derived from its observation) is a *guess*. Accordingly, **there is no such thing as certain belief** pertaining to the physical universe; all belief pertaining to the physical universe[50] is *uncertain*.

This unfortunate but nevertheless unavoidable reality makes reasoning about the deepest subtleties of the physical universe distinctly counter-intuitive. In the last ~370 years in particular, a number of mechanisms have been proposed for **reasoning about uncertainty mathematically**, including:

- Dempster-Shafer belief functions[51]
- possibility measures[52]
- probability measures[53]
- ranking functions[52]
- relative likelihood[52]
- plausibility measures[52] (note that plausibility measures generalise of all of the above!)

Of these, **probability theory** is by far the most widely used in statistics, science, and engineering:

- a **probability** is a number between 0 and 1 expressing (in the modern Bayesian interpretation) "strength of belief", i.e. the degree to which an assertion is supported by the available evidence
- probability theory has been **axiomatised**, notably by Kolmogorov (1933); such axiomatisations describe the rules via which probabilities may be manipulated (i.e. reasoned about).

The cognitive primitives induction, deduction, and abduction may be applied **probabilistically**:

- **induction**:
  - rather than a set of observations *definitely* containing pattern X, if the data are noisy then one might conclude that they *might* contain pattern X (with a certain probability)
  - similarly, having identified a specific pattern X as possibly being contained in previous observations, a new observation *might* match that pattern (with a certain probability)
- **deduction –** rather than concluding (for example) that "it will definitely rain today", by applying the probability axioms one might conclude that "there's a 70% chance that it will rain today"
- **abduction –** similarly, on learning that "Socrates is mortal", one might conclude (based on past experience) that "there's a 90% chance that Socrates is a man, and a 10% chance he is not".

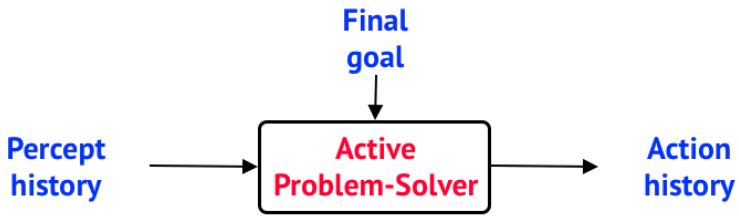
Note that, in practice, probabilities themselves are usually estimates, i.e. *guesses!*

[50] such as whether or not the physical universe exists, whether or not you exist, who gave birth to you, or what time it is
[51] Shafer (1976); Yager et al. (2004); Halpern (2017)
[52] Halpern (2017)
[53] Kolmogorov (1933); Jeffreys (1939); Cox (1946); Pearl (1988) Bernardo and Smith (2000); Jaynes (2003); Jeffrey (2004); Benaroya and Han (2005); Rosenthal (2006); Stone (2013); Venkatesh (2013); Gelman et al. (2014); Halpern (2017)

2.2.29 ACTIVE PROBLEM-SOLVERS

So far we have been discussing *passive* problem-solvers:

An **active problem-solver** (or **agent**) takes as input a *continuous stream of percepts* (its *percept history*), and performs a *continuous stream of actions* (its *action history*) as output, all in pursuance of its (fixed) *final goal* (the invariant condition that the agent strives to continuously maintain):



where:

- a **percept** is a digitised *input* received from a *sensor* (such as a video camera, or microphone)
- an **action** is a digitised *output* sent to an *effector* (such as an actuator, display screen, or speaker)
- the **final goal** might be e.g.:
  - "Do nothing"
  - "Do something"
  - "Calculate the decimal expansion of pi"
  - "Maximise the total number of paperclips that exist"
  - "Maximise shareholder value for XYZ Corp"
  - "Maximise GDP for country X"
  - "Maximise human happiness"
- ... all of which are problematic in one way or another.

We shall return to the problem of formulating a final goal in Section **??**.

Note that:

- humans, organisations, and robots may all be viewed as active problem-solvers, i.e. agents
- whereas a *passive* problem-solver must attempt to solve a specified problem $\langle P, Q, R \rangle$ using just the information that it already has and the physical resources R specified by the problem statement, an *active* problem-solver may acquire *additional information* and *additional resources* through deliberate (i.e. **planned**[54]) interaction with the physical universe.

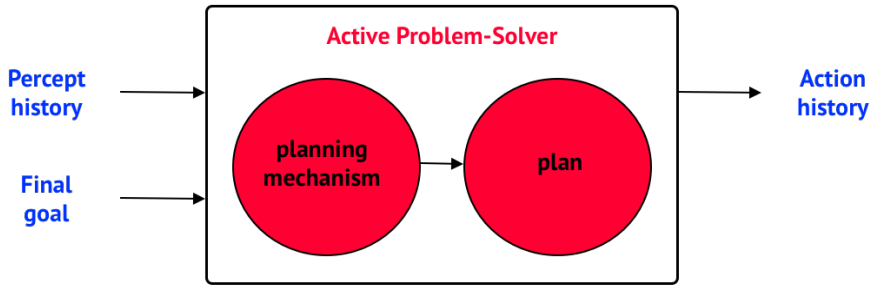
[54] see Section 2.2.30

2.2.30 CONTINUOUS PLANNING

We define:

- a **planning mechanism** to be an algorithm which, when executed, generates a **plan**
- a **plan** to be an algorithm which, when executed, generates a sequence of **actions**[55].

Any active problem-solver must necessarily generate actions *somehow*, even if only one at a time. Accordingly, if we look inside an active problem-solver, we will necessarily see some kind of planning mechanism [Allen, Hendler, and Tate (1990); Ghallab, Nau, and Traverso (2016)].



The planning mechanism strives to solve (and continuously re-solve) the following problem: "Given the current state of the (assumed) physical universe (as indicated by the percept history), formulate a plan which, when executed, will generate a sequence of actions (to be appended to the action history) whose most likely **causal effect**[56] is to maximise progress towards (and ideally achieve) the final goal". Append to this an English description of a specific final goal, and the result is an English specification (meta-level problem statement) for the body of the planning loop that might be given to a human programmer, or to a human AI designer (meta-level passive problem-solver).

Accordingly, following Section 2.2.24, the body of the planning loop is effectively an object-level passive problem-solver to be designed by a meta-level passive problem-solver (human AI designer)[57].

There are many ways in which the AI designer might choose to solve this problem. For example, it might be possible, in some cases, to reformulate the final goal as a finite set of *production rules*[58] to be blindly followed in response to incoming percepts, thereby generating a sequence of actions "whose most likely causal effect is to make progress towards (and ideally achieve) the final goal". In other words, rather than the final goal being an explicitly-represented component of the planning mechanism, the desired behaviour is instead an *emergent property* of the set of production rules [Kowalski (2011b)]. Should the production rules include any pre-calculated information derived (by the AI designer) from the set of all possible object-statements then the body of such a planning loop might not be maximally-intrinsic. As argued in Section 2.2.26, maximally-intrinsic solutions facilitate the path to superintelligence, and (if that is the objective) are therefore greatly preferable.

We will say that an active problem-solver is **maximally-intrinsic** if every passive problem-solver that it incorporates (for example, as the body of the planning loop) is maximally-intrinsic.

[55] in the simplest possible case, a planning mechanism might generate plans which generate a single action at a time

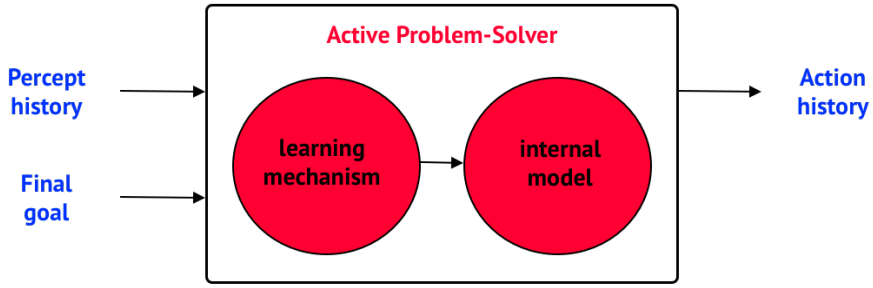
[56] see Section 2.2.37

[57] we shall henceforth use the designation "AGI X" to refer to either (a) a *passive* problem-solver at level AGI X, or (b) an *active* problem-solver *incorporating* a passive problem-solver at level AGI X (e.g. as the body of the planning loop)

[58] a production rule has the general form: if *this happens* then *do this*

2.2.31 CONTINUOUS LEARNING

If we look inside an active problem-solver, we might also see some kind of **learning mechanism**:



A learning mechanism strives to solve (and continuously re-solve) the following problem: "Given the current percept history, construct an **internal model**[59] capturing (as accurately as possible) the **structure**[60] of the (assumed) physical universe, including all of its nuanced complexity[61]". Due to the importance of information to problem-solving, an active problem-solver that incorporates a continuous learning mechanism will be able to apply its accumulated **learned knowledge** to future problems (potentially finding better solutions as a result), whereas an active problem-solver that does not incorporate such a learning mechanism will be unable to grow with experience in the same way[62].

Again following Section 2.2.24, the body of the learning loop is effectively an object-level passive problem-solver to be designed by a meta-level passive problem-solver (e.g. a human AI designer).

As before, there are many ways in which the AI designer might choose to solve this problem [Michalski, Carbonell, and Mitchell (1983); Mitchell (1997); Jain (1999); MacKay (2003); Bishop (2006); Rasmussen and Williams (2006); Murphy (2012); Shalev-Shwartz and Ben-David (2014); Goodfellow, Bengio, and Courville (2016); Sutton and Barto (2018); Faul (2020); Schulte (2023)]. Some solutions will be maximally-intrinsic, whereas others will not. Maximally-intrinsic solutions facilitate the path to superintelligence, and (if that is the objective) are therefore greatly preferable.

Clearly, an active problem-solver will only qualify as **maximally-intrinsic** if the body of its planning loop *and* the body of its learning loop (assuming that it has one) are maximally-intrinsic.

---

[59] see Section 2.2.34

[60] see Section 2.2.33

[61] including, but not limited to [Adler (1974)]: Matter and Energy (Atoms; Energy, Radiation, and States of Matter; the Universe), The Earth (Earth's Properties, Structure, Composition; Earth's Envelope; Surface Features; Earth's History), Life (The Nature and Diversity of Life; The Molecular Basis of Life; The Structures and Functions of Organisms; The Behaviour of Organisms; The Biosphere), Human Life (The Development of Human Life; The Human Body: Health and Disease; Human Behaviour and Experience), Society (Social Groups: Ethnic Groups and Cultures (Peoples and Cultures of the World; The Development of Human Culture; Major Cultural Components and Institutions of Societies; Language and Communication); Social Organisation and Social Change; The Production, Distribution, and Utilization of Weath; Politics and Government; Law; Education), Art (Art in General; Particular Arts), Technology (Nature and Development of Technology; Elements of Technology; Fields of Technology), Religion (Religion in General; Particular Religions), History (Ancient Southwest Asia, North Africa, and Europe; Medieval Southwest Asia, North Africa, and Europe; East, Central, South, and Southeast Asia; Sub-Saharan Africa to 1885; Pre-Columbian America; The Modern World to 1920; The World Since 1920), and Branches of Knowledge (Logic; Mathematics; Science; History and the Humanities; Philosophy; Preservation of Knowledge) — note that **natural languages such as English are part of the structure of the universe**; a learning mechanism should be able to assimilate *all* natural languages along with the rest of the structure of the universe!

[62] this quality of *adaptability* is often cited in the literature as being a core principle of intelligence [see e.g. Wang (2013)]

2.2.32 THE SENSE-EFFECT LOOP

We now have enough pieces of the active-problem-solver puzzle to see how they all fit together:

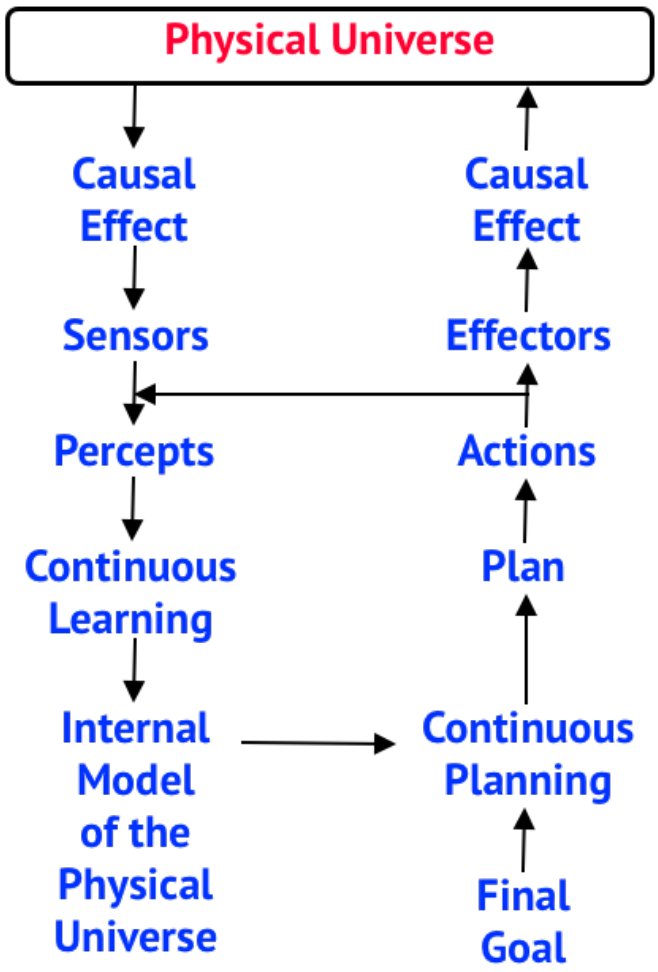


Figure 2: The active problem-solver **sense-effect loop**

We shall unpack Figure 2 over Sections 2.2.33 to 2.2.38, starting with the physical universe.

2.2.33 THE STRUCTURE OF THE PHYSICAL UNIVERSE

Let's assume for a moment that the physical universe actually exists. What do we know about it...?

We know that it's not *completely homogeneous*. If it were, wherever we looked, with whatever sensor, at any time, we would see the same thing. But that's not what we see. We also know that it's not *completely random*. If it were, wherever we looked, with whatever sensor, at any time, all we would see is random data (like an old TV screen with no signal). But that's not what we see either.

What we see instead is something in between *complete homogeneity* and *complete randomness*. In other words, **the universe has structure**, and that structure is *noisy*, but not *completely* noisy.

2.2.34 INTERNAL MODELS OF THE PHYSICAL UNIVERSE, AND DEPTH OF UNDERSTANDING

Following Figure 2, the physical universe **causally affects** the **sensors** (Section 2.2.29) of an active problem-solver, or agent, *G*, thereby generating a continuous stream of **percepts**. These percepts then become the primary input into *G*'s **continuous learning mechanism** (Section 2.2.31), the body of which comprises a passive problem-solver (Section 2.2.7) utilising various cognitive primitives such as (for example) strong probabilistic induction, deduction, and abduction (Sections 2.2.27 and 2.2.28). As its primary output, the continuous learning mechanism maintains an **internal model** capturing (as accurately as possible) the **structure** (Section 2.2.33) of the physical universe[63]; e.g.[64]:

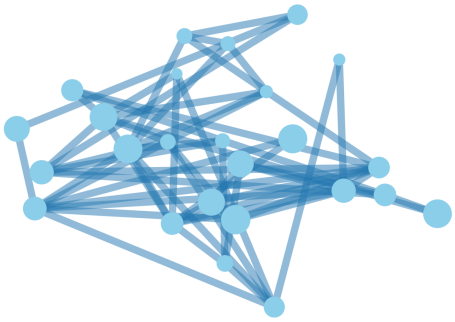


Figure 3: **Shallow** internal model (and therefore understanding) of the physical universe

Let's imagine that Figure 3 is an internal model of the physical universe, where each circle represents some *feature* (corresponding to an identifiable part of the structure of the physical universe) and each line represents a *relationship* between two or more features. Let's further imagine that Figure 3 captures only a fragment of the structure of the *actual* physical universe — in other words, there are lots of *missing features* and *missing relationships* that are not captured by the model. In such a scenario, we would say that agent *G* has only a **shallow understanding** of the physical universe.

Let's imagine that agent *G* observes the physical universe for (say) another *ten years*, and, given these additional percepts, *G*'s learning mechanism is able to construct a more detailed internal model:

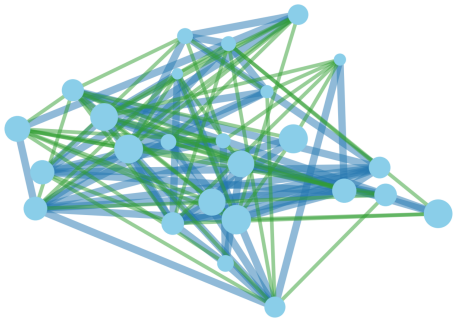


Figure 4: **Intermediate** internal model (and therefore understanding) of the physical universe

[63] note that if agent *G* lacks a continuous learning mechanism then its internal model of the physical universe will be fixed
[64] the physical universe is insanely complex, and so (potentially) is any internal model of it; we will try to keep things simple!

Finally, let's imagine that agent *G* observes the physical universe for a further *hundred years*, thereby allowing *G*'s learning mechanism to construct a much more detailed internal model:

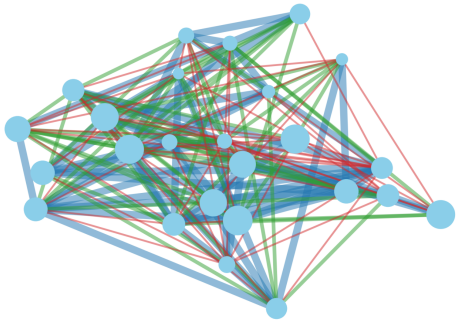


Figure 5: **Deep** internal model (and therefore understanding) of the physical universe

In such a scenario, we would say that agent *G* has a **deep understanding** of the physical universe.

As already alluded in Section 2.2.26, "a **deep understanding** of any subject [such as the physical universe] provides much more **problem-solving information** than a **shallow understanding**".

2.2.35 INDEPENDENCE

It is instructive to consider the various relationships that might (or might not) exist between features of the physical universe, and which might (or might not) be captured by an internal model of it.

The simplest case is when **no relationship exists** at all. Thus two features *x* and *y* are said to be **independent** when information about one gives no information about the other, for example:

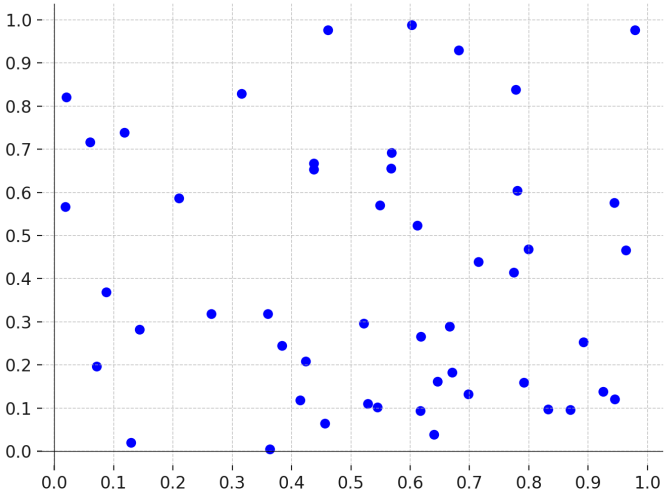
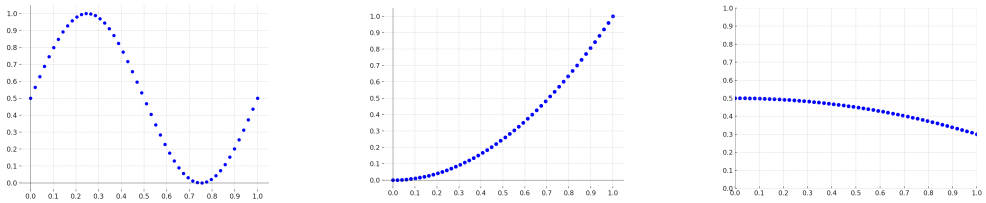


Figure 6: Independent features *x* and *y*

2.2.36 CODEPENDENCE

Alternatively, features x and y may be **codependent**, for example:

- **association**: if either feature changes, the other also changes[65]
- **+ve correlation**: if either feature changes, the other changes *in the same direction*
- **–ve correlation**: if either feature changes, the other changes *in the opposite direction*:

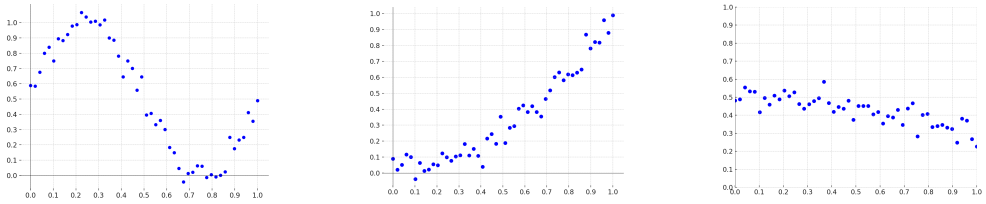


(a) Association (b) Positive correlation (c) Negative correlation

Figure 7: Examples of simple codependence

As with induction, deduction, and abduction, the above relationships may be *probabilistic* (noisy):

- **association**: if either feature changes, the other *tends* to change
- **+ve correlation**: if either feature changes, the other *tends* to change in the same direction
- **–ve correlation**: if either feature changes, the other *tends* to change in the opposite direction:



(a) Noisy association (b) Noisy positive correlation (c) Noisy negative correlation

Figure 8: Examples of noisy codependence

Note that, if the relationship between two simply codependent features is known, then one may be **calculated** (i.e. **exactly predicted**) from the other. Similarly, if the relationship between two noisily codependent features is known, then one may be **estimated** (i.e. **approximately predicted**) from the other. Accordingly, whenever codependent relationships exist between features of the physical universe, and these relationships are captured by agent G's internal model of the physical universe, then G will be able to use its **internal model** to make **predictions** about the **actual physical universe**.

[65] but not necessarily in the same direction, and not necessarily in the opposite direction either!

2.2.37 CAUSALITY

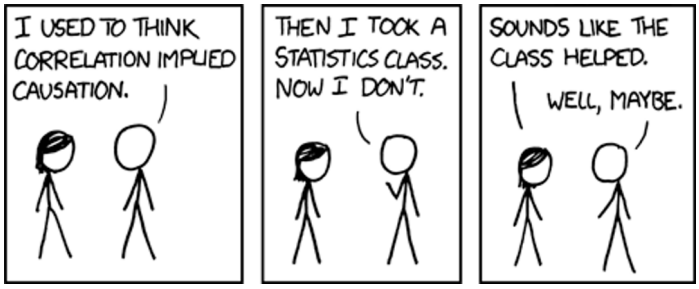
Consider the following statement (which, for our present purposes, we shall assume to be true):

In the English seaside town of Brighton, ice cream sales are positively correlated with drownings.

As already alluded in Section 2.2.36, this means that, if the numerical relationship between "daily ice cream sales" and "daily drownings" is approximately known, then, given one, we may confidently estimate the other. But does this mean that ice cream sales **cause** drownings? Because, if that were the case, restricting ice cream sales would lead to a decrease in the number of drowning deaths!

Clearly, ice cream sales do not cause drownings. A much more plausible explanation is that some common factor, such as hot weather, causes both an increase in ice cream sales and an increase in the number of people taking a swim (some ~fixed proportion of whom will unfortunately drown).

This is an example of the well-known phenomenon that **correlation does not imply causality**[66]. This has important implications for active problem-solving, because it means that, when constructing a **plan of actions**, i.e. when trying to predict the **causal effect** of each **candidate action**, an agent *G* cannot simply rely on any **correlations** that might be present in its internal model of the physical universe. Instead, in order for *G* (when formulating a plan) to be able to predict the **likely causal effect** of each candidate action, *G*'s **internal model** of the physical universe must include relationships between features that capture the **cause-and-effect behaviour** of the **actual** physical universe.



2.2.38 FROM INTERNAL MODEL AND FINAL GOAL TO ACTIONS AND CAUSAL EFFECTS

Returning to Figure 2, the **internal model** maintained by *G*'s **continuous learning mechanism** (Section 2.2.31) then becomes — along with *G*'s **final goal** (Section 2.2.29) — the primary input into *G*'s **continuous planning mechanism** (Section 2.2.30), the body of which (exactly analogously to *G*'s continuous learning mechanism) comprises a passive problem-solver (Section 2.2.7) utilising various cognitive primitives such as (for example) strong probabilistic induction, deduction, and abduction (Sections 2.2.27 and 2.2.28). As its primary output, and guided by the cause-and-effect relationships in its internal model of the physical universe, the continuous planning mechanism formulates a **plan** whose execution generates a continuous stream of **actions**, which, via *G*'s **effectors** (Section 2.2.29) then **causally affect** the physical universe[67]. If the causal effects of the actions generated by the plan serve to maximise progress towards (and ideally achieve) the final goal, then the plan formulated by the continuous planning mechanism may be said to be **causally effective**.

[66] sometimes also called *causation* [Pearl (2009); Halpern (2016); Pearl and Mackenzie (2019)]

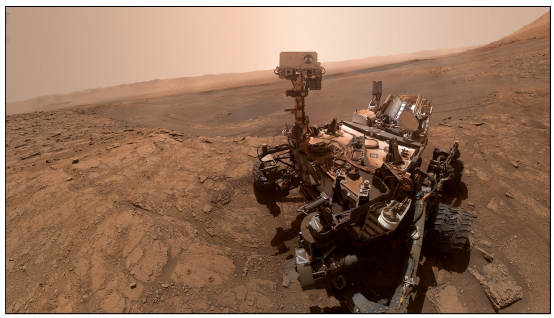
[67] note how *actions* are also **looped back** as *percepts*; this not only maintains an automatic record (in *G*'s percept history) of every action sent to the effectors (over which *G* may then perform **induction**), it also allows *G* to **imagine** its own percepts

2.2.39 AUTONOMOUS ROBOTS

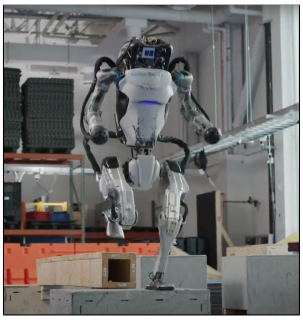
Being a physical system (ultimately comprising computer hardware etc), an active problem-solver must necessarily possess some kind of physical **body**, including its physical sensors and effectors.

In the limit, an active problem-solver may be mounted on a **physical frame**, such that its sensors and effectors are sufficient for **locomotion** (i.e. the ability to move around in physical space).

For example, NASA's *Curiosity* Mars rover (see Figure 9a) is capable of surface locomotion.



(a) *Curiosity* takes a selfie



(b) *Atlas* goes for a run

Figure 9: Examples of contemporary robots capable of surface locomotion

Although the *Curiosity* rover is able to do some things "all by itself" — such as navigate from point A to point B, and select interesting rocks to sample — it is still nevertheless mostly controlled by NASA engineers on Earth. It is therefore only **partially-autonomous**, rather than **fully-autonomous**.

In order for a robot [Corke (2011)] to be genuinely fully-autonomous, the body of its planning loop must comprise a sufficiently performant object-level passive problem-solver that the robot is able to perform its desired function (i.e. pursuit of its final goal) **without any human supervision**.

In the popular imagination, the **ultimate fully-autonomous robot** (UFAR) comprises:

1. a **humanoid frame** with sensors[68] and effectors[69] (e.g. c/o Boston Dynamics — Figure 9b)
2. a **maximally-intrinsic active problem-solver**[70] (Section 2.2.29):
   (a) incorporating:
      (i) a **continuous planning loop** (Section 2.2.30)
         — whose body comprises a **superintelligent passive problem-solver** (AGI O/Z)
      (ii) a **continuous learning loop** (Section 2.2.31)
         — whose body comprises a **superintelligent passive problem-solver** (AGI O/Z)
   (b) that is **maximally-aligned with human beings in perpetuity**.

We shall explore the concept of **maximal alignment** over Sections 2.2.40 to 2.2.49.

[68] e.g. vision (visible light, thermographic, LIDAR, radar), hearing (audio, ultrasonic, sonar), touch (tactile/haptic), smell/taste (olfactory), position (shaft encoders, magnetometer, gyroscope, accelerometer, clock, GPS), incoming cellular/WiFi

[69] e.g. graphical displays, speakers, locomotors (legs, wheels), manipulators (arms, hands), outgoing cellular/WiFi

[70] arranged such that UFAR's sensors input to the percept history, and the action history outputs to UFAR's effectors

2.2.40 LIVENESS AND SAFETY

We hereby extend the concept of **agent** to include the following, viewed as active problem-solvers:

- individual humans
- organisations
- autonomous robots.

The following are **desirable properties** of (the behaviour of) any agent G:

1. **good (i.e. desirable) things happen**[71] — usually referred to as **liveness**[72]
2. **bad (i.e. undesirable) things don't happen** — usually referred to as **safety**.

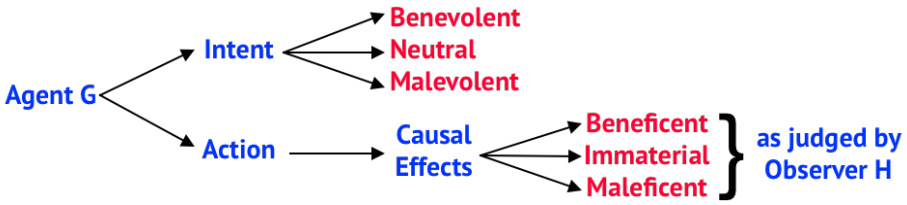
Things are a little more complicated than might appear at first glance because:

- the concepts of **desirable** and **undesirable** are **subjective** (i.e. **relative to some observer**)
- **intent** and **effect** (i.e. causal effect, as described in Section 2.2.37) are **two different things**.

The precise terminology pertaining to **safety** and **liveness** might be summarised as follows:

benevolent	(adjective)	(an agent G) intending to have desirable causal effect (as judged by observer H)
neutral	(adjective)	(an agent G) lacking any intent in respect of causal effect (as judged by observer H)
malevolent	(adjective)	(an agent G) intending to have undesirable causal effect (as judged by observer H)
beneficent	(adjective)	(an agent G, or action) having desirable causal effect (as judged by observer H)
immaterial	(adjective)	(an agent G, or action) not having any causal effect (as judged by observer H)
maleficent	(adjective)	(an agent G, or action) having undesirable causal effect (as judged by observer H)
benevolence	(noun)	the quality of intending to have desirable causal effect (as judged by observer H)
neutrality	(noun)	the quality of lacking any intent in respect of causal effect (as judged by observer H)
malevolence	(noun)	the quality of intending to have undesirable causal effect (as judged by observer H)
beneficence	(noun)	the quality of having desirable causal effect (as judged by observer H)
immateriality	(noun)	the quality of not having any causal effect (as judged by observer H)
maleficence	(noun)	the quality of having undesirable causal effect (as judged by observer H)

It's possible for an agent G (e.g. a human, organisation, or autonomous robot) to be **benevolent** (i.e. to have benevolent **intent**) and yet for the **causal effects** of G's actions to nevertheless be judged (by some observer H) as being *some combination* of **beneficent**, **immaterial**, and **maleficent**.



Note that different observers H_1 and H_2 might judge the causal effects of G's actions differently.

[71] as a result of G's actions

[72] sometimes also called **progress** — an agent that doesn't do anything may be perfectly safe but, well, it doesn't *do* anything!

2.2.41 ALIGNMENT

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere once we have started it ... then we had better be quite sure that the purpose put into the machine is the purpose which we really desire and not merely a colorful imitation of it. Wiener (1960)

Simply stated, **alignment = liveness + safety**, i.e. **good things happen, and bad things don't**.

As already alluded, the concepts of *good* (desirable) and *bad* (undesirable) are subjective, i.e. only meaningful relative to some **observer** H, or, in general, some **population** of observers $\mathcal{H}$[73].

An agent G is **aligned with human values** $\mathcal{V}$[74] if G's behaviour is consistent with $\mathcal{V}$.

2.2.42 HUMAN VALUES

If G is a purposefully-engineered maximally-intrinsic agent (such as an organisation, or an autonomous robot) which we desire to be **aligned with human values** $\mathcal{V}$, then G must have **knowledge of human values** $\mathcal{V}$ in order to be able to behave accordingly. There are two ways in which this may be achieved:

(1) we (as G's designers) work out what human values $\mathcal{V}$ to use, and **code them** into G[75], or

(2) by **carefully observing humans**, agent G works out for itself what human values $\mathcal{V}$ to use.

Despite many centuries of effort by mankind's greatest moral philosophers, there does not seem to be any universally agreed upon set of human values $\mathcal{V}$ that we can simply code into G. On the contrary, we (humans) can't even agree on when it is, and isn't, OK to take a human life. Instead, human values $\mathcal{V}$ vary from individual to individual, from culture to culture, and even across time. What were once acceptable human values 300 years ago are very different from what is considered acceptable today, and doubtless many of today's human values will seem barbaric 300 years from now. As responsible AI designers, we must be careful not to impose *our* values onto future humans!

Thus option (1) doesn't seem viable, which leaves option (2) whereby, in order to align itself with humans, agent G carefully observes humans and, from those observations, determines (or at least estimates as accurately as possible) an appropriate set of human values $\mathcal{V}$. A great advantage of this approach is that agent G **continually re-aligns itself in perpetuity** as human values evolve over time.

Figure 10: Human values (as imagined by Midjourney)

[73] assumed, unless explicitly stated otherwise, to be the population of all humans (living and future)
[74] corresponding to what humans (on aggregate) count as *good* (desirable) and *bad* (undesirable)
[75] for example, via G's constitution (if G is an organisation), or G's final goal (if G is an autonomous robot)

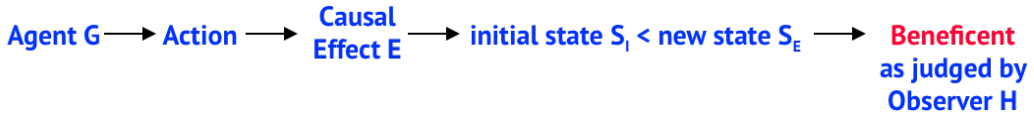
2.2.43 ORDINAL PREFERENCES

The concept of "human values $\mathcal{V}$" is a little vague. It's much easier to think in terms of **preferences**.

In general, an agent G performs a sequence of actions in pursuit of its goals. Any such **action sequence** $\mathcal{A}$ may have **possible causal effects** $E_1, ..., E_n$ with corresponding **probabilities** $P_1, ..., P_n$[76].

Imagine $\mathbb{States}$, the set of all possible states of the physical universe $\{S_1, ..., S_\omega\}$. Given sufficient information, $\mathbb{States}$ may be strictly ordered such that $S_x < S_y$ if and only if observer H judges state S_x to be **less desirable** than state S_y. Such an ordering is said to express H's **individual preferences** [Arrow (1951); Hansson and Grüne-Yanoff (2022)]. If the initial state is S_I then causal effect E corresponding to new state S_E is **beneficent**[77] if $S_I < S_E$, and **maleficent**[77] if $S_E < S_I$. If states are partially ordered then some causal effects may be **immaterial**[77], i.e. neither beneficent nor maleficent:

Beneficent causal effect E:



Immaterial causal effect E:

Maleficent causal effect E:

Note that, in general, observer H's preferences are not fixed, and may change over time.

2.2.44 UTILITY FUNCTIONS, AND EXPECTED UTILITY

Some preferences are more strongly held than others. A **utility function** U for observer H assigns a numeric value ≥ 0 to each state. If the initial state is S_I then causal effect E corresponding to new state S_E is **beneficent** [with **impact** $U(S_E) - U(S_I)$] if $U(S_I) < U(S_E)$, **maleficent** [with **impact** $U(S_I) - U(S_E)$] if $U(S_E) < U(S_I)$, and **immaterial** [with **impact** 0] if $U(S_E) = U(S_I)$. The **expected utility** (according to U) of an action sequence $\mathcal{A}$ having possible causal effects $E_1, ..., E_n$ (corresponding to new states $S_1, ..., S_n$) with corresponding probabilities $P_1, ..., P_n$ may be calculated by multiplying, for each possible causal effect E_i, the utility (according to U) of the corresponding new state S_i by the probability P_i that the causal effect in question will occur, and then summing the results:

$$\textbf{expected utility} = [U(S_1) \times P_1] + [U(S_2) \times P_2] + [U(S_3) \times P_3] + ... + [U(S_n) \times P_n].$$

[76] in other words, P_i is the probability (between 0 and 1) that causal effect E_i will occur as a result of action sequence $\mathcal{A}$
[77] as judged by H

2.2.45 ELICITING HUMAN PREFERENCES

Captain Renault: And what in heaven's name brought you to Casablanca?
Rick Blaine: My health. I came to Casablanca for the waters.
Captain Renault: The waters? What waters? We're in the desert.
Rick Blaine: I was misinformed.

Casablanca, 1942

Let's imagine that we desire agent G to align its behaviour with an **individual human** H. One approach would be for agent G to carefully observe H and, from those observations, determine (or at least estimate as accurately as possible) what H's preferences are (expressed as a utility function $\hat{U}$) in order that G may then behave in a manner that is consistent with those (estimated) preferences.

In practice, **preference elicitation** is surprisingly difficult. For example:

- Being human, H may sometimes be less than entirely **rational** (this being the kindest way we can find to say it!) It will rarely be in H's best interest for G to realise an irrational preference.
- H may sometimes be less than entirely **truthful** (e.g. in embarrassing social contexts).
- H might not always **know what they want** (either precisely, or even at all).
- Even if they do, H might not always **express their preferences** clearly and unambiguously[78].
- Accordingly, H's **stated preferences** are not necessarily the same as H's **actual preferences**[79].
- It's possible that H is less than entirely **well-informed**[80] to some significant degree in respect of the preferability of some future state, perhaps as a result of being deliberately **misinformed** by another party [Davies (2009); Oborne (2021)]. For example, H might request a rat poison cupcake for dessert, on the basis of what somebody had told them, or some YouTube video that they had seen. It will rarely be in H's best interest for G to realise an ill-informed preference.
- It's possible that an actor J has **manipulated** H into ostensibly having some preference P to a greater or lesser degree than H would have had had it not been for J's manipulation [Handelman (2009); Coons and Weber (2014); Noggle (2022)]. In such situations, H's apparent preference P is less than entirely genuine, having been influenced by J's manipulation, and will in some cases effectively be J's preference rather than H's. It will not necessarily be in H's best interest for G to realise a preference P that has not been determined entirely freely by H.

Accordingly, H's best interests would be much better served if agent G were to carefully observe H and, from those observations, determine (or at least estimate as accurately as possible)[81] what H's actual rational well-informed freely-determined preferences (expressed as a utility function $\hat{U}$) **would be** if H were entirely rational, well-informed, and free from manipulation by any party[82].

[78] for example, H might be a pre-verbal human child
[79] just ask *The Sorcerer's Apprentice* [Wiener (1960)] or *King Midas* [Russell (2019)]
[80] here, a *well-informed* person does not simply have *access* to the relevant information, but also fully *understands* it
[81] e.g. via an arbitrarily complex combination of strong probabilistic induction, deduction, and abduction (2.2.27 and 2.2.28)
[82] such preferences are often called **idealised preferences** [see e.g. Hendrycks (2023)]

2.2.46 AGGREGATING HUMAN PREFERENCES

In order for agent G to align itself with **all humans** $\mathcal{H}$, one approach would be for G to:

(1) carefully observe **each individual human** H_i in the population of all humans $\mathcal{H}$[83] and, from those observations, determine (or at least estimate as accurately as possible) what H_i's actual rational well-informed freely-determined preferences (expressed as a utility function $\hat{U}_i$) would be if H_i were entirely rational, well-informed, and free from manipulation by any party

(2) calculate, in some "perfectly fair" manner, an **aggregated utility function** $\hat{U}$ for the population of all humans $\mathcal{H}$ from the **individual utility functions** $\hat{U}_i$ for each individual human H_i.

Although many different methods have been proposed for aggregating individual utility functions [Arrow (1951); Gabriel (2020); List (2022)], none is ideal[84]. In other words, there is no unambiguously "perfectly fair" method of doing so[85] — which effectively makes how to do so a **value judgement**.

One possible solution therefore would be to split step (2) above into two parts, as follows:

(2a) calculate, in some sensible manner, an interim aggregated utility function $\bar{U}$ for the population of all humans $\mathcal{H}$ from the individual utility functions $\hat{U}_i$ for each individual human H_i

(2b) calculate, in some manner consistent with $\bar{U}$, an aggregated utility function $\hat{U}$ for the population of all humans $\mathcal{H}$ from the individual utility functions $\hat{U}_i$ for each individual human H_i.

In other words, the interim aggregated utility function $\bar{U}$ for all humans calculated at step (2a) constrains how the final aggregated utility function $\hat{U}$ for all humans is calculated at step (2b)[86].

2.2.47 REALISING HUMAN PREFERENCES

Once agent G has estimated an aggregated utility function $\hat{U}$ for the population of all humans $\mathcal{H}$, G may proceed to behave (perform a sequence of actions) consistent with $\hat{U}$, for example as follows:

- G cannot possibly consider every possible sequence of future actions, as that would be infinite
- instead, G must necessarily look ahead by some finite amount, say the next $M \geq 1$ actions
- using its **internal model** (Section 2.2.34) as maintained by its learning mechanism (Section 2.2.31), G attempts to **predict**, for every possible action sequence $\mathcal{A}$ of length M:
  - the possible **causal effects** (of action sequence $\mathcal{A}$) $E_1, ..., E_n$
  - their corresponding **probabilities** $P_1, ..., P_n$
- G then performs the action sequence yielding the greatest **expected utility** according to $\hat{U}$.

Thus agent G strives to behave in such a way as to **maximise expected utility** for all humans $\mathcal{H}$.

[83] or at least a statistically meaningful sample of $\mathcal{H}$ — the larger the better

[84] one problem, for example, is that individual utilities are not easily comparable, i.e. "10 units of utility" for one person does not necessarily equal "10 units of utility" for another — *inter-person utility calibration* requires additional information

[85] which means that the best that an agent G can ever hope to achieve in actual practice is to try to find some (not necessarily **perfectly fair** but nevertheless) **maximally fair** way in which to calculate the estimated aggregated utility function $\hat{U}$

[86] the process of constructing an aggregated utility function $\hat{U}$ could of course be extended to more than two levels

2.2.48 PERFECT ALIGNMENT

We will say that agent G is **perfectly-aligned** with population $\mathcal{H}$ (such as the population of all humans, both living and future) whose aggregated preferences are encapsulated by utility function U if:

1. the total impact of the **beneficent causal effects** of G's actions are **maximised**
2. the total impact of the **maleficent causal effects** of G's actions are **minimised**[87].

In other words, good things (resulting from G actions) are maximised, and bad things (resulting from G's actions) are minimised, where what counts as good and bad, as well as the trade-offs between good and bad things, are determined by utility function U (encapsulating $\mathcal{H}$'s aggregated preferences).

Unfortunately, **perfect alignment is all-but-impossible in practice**. For example, in general:

– G will only ever have partial information (about the physical universe, humans, etc)
– G will never have enough time and other physical resources (particularly compute)
– G's estimation $\hat{U}$ of $\mathcal{H}$'s aggregated utility function U will never be entirely accurate
– G's predictions pertaining to the possible causal effects of its actions will never be entirely accurate.

2.2.49 MAXIMAL ALIGNMENT

We will say that agent G is **maximally-aligned** with population $\mathcal{H}$ (such as the population of all humans, both living and future) whose aggregated preferences are encapsulated by utility function U if the **distance from perfect alignment** to the degree to which agent G is aligned is **minimised** (consistent with U)[88]. We make the following observations:

- **maximal alignment requires considerable knowledge and understanding** — in order to achieve maximal alignment in respect of the population of all humans $\mathcal{H}$, a maximally-intrinsic agent G must necessarily gain a deep understanding of a wide range of complex and nuanced subjects, including humans, human languages[89], mathematics (including uncertainty), and physics (including causality); G will also need to maintain a detailed and accurate internal model of the physical universe in order to be able to make accurate predictions about it
- **maximal alignment requires considerable problem-solving ability** — in order to reason about all of the above, G will need to apply robust critical thought and problem-solving (e.g. strong probabilistic induction, deduction, and abduction) to complex and nuanced concepts
- **maximal alignment requires considerable compute** — solving problems pertaining to all of humanity will require massive compute[90]; without it, maximal alignment is compromised
- **most AGI designs will be misaligned by default** — maximal alignment requires highly-specific technical capabilities which the vast majority of AGI designs will simply not possess.

[87] analogously, *perfect misalignment* = total beneficent impact is minimised and total maleficent impact is maximised
[88] analogously, *maximal misalignment* = the distance from perfect misalignment is minimised (consistent with $-U$)
[89] 35 languages cover 95% of all humans: English, Mandarin Chinese (inc Standard Chinese), Hindi, Spanish, French, Modern Standard Arabic, Bengali, Portuguese, Russian, Urdu, Indonesian, Standard German, Japanese, Nigerian Pidgin, Egyptian Arabic, Marathi, Telugu, Turkish, Tamil, Yue Chinese (inc Cantonese), Vietnamese, Wu Chinese (inc Shanghainese), Tagalog, Korean, Iranian Persian, Hausa, Swahili, Javanese, Italian, Punjabi, Gujarati, Thai, Kannada, Amharic, Bhojpuri
[90] we suspect easily > 10 orders of magnitude more compute per USD (and ideally per W) than is available in 2024

2.2.50 ALIGNMENT IS EVERYTHING!

Compare the following best- and worst-case scenarios:

- **best-case scenario** — imagine that agent G is an AGI Z (maximally-super-inventive + maximally-super-knowledgeable + maximally-super-resourced, i.e. the most intelligent superintelligent AGI that it's possible to build)[91], as well as maximally-*aligned* with the population of all humans $\mathcal{H}$ whose aggregated preferences are encapsulated by utility function U; given all of the above, G will strive to behave in such a way as to *maximise* expected utility for all humans $\mathcal{H}$ to the maximum extent that is possible in actual practice; in other words, good things (resulting from G actions) are *maximised*, and bad things (resulting from G's actions) are *minimised*, where what counts as good and bad, as well as the trade-offs between good and bad things, are determined by utility function U (encapsulating $\mathcal{H}$'s aggregated preferences); or, to put it yet another way, G will strive to create *the best possible utopia* for all mankind

Figure 11: A near-utopia for all mankind for all eternity (as imagined by Midjourney)

- **worst-case scenario** — imagine that agent G is the exact same AGI Z as above except that its final goal has been modified such that G is now maximally-*misaligned* with the population of all humans $\mathcal{H}$ whose aggregated preferences are encapsulated by utility function U; given all of the above, G will strive to behave in such a way as to *minimise* expected utility for all humans $\mathcal{H}$ to the maximum extent that is possible in actual practice; in other words, good things (resulting from G actions) are *minimised*, and bad things (resulting from G's actions) are *maximised*, where what counts as good and bad, as well as the trade-offs between good and bad things, are determined by utility function U (encapsulating $\mathcal{H}$'s aggregated preferences); or, to put it yet another way, G will strive to create *the worst possible dystopia* for all mankind.

[91] consistent with aggregated human preferences — e.g. not so much compute that Earth's atmosphere boils away!

Figure 12: A near-dystopia for all mankind for all eternity (as imagined by Midjourney)

The only difference between these two scenarios is a **minus sign** — in the best-case scenario agent *G* is maximally-*aligned*, and in the worst-case scenario agent *G* is maximally-*misaligned*. **Every intermediate scenario (imaginable or not) in between the best and worst cases is also possible**.

2.2.51 X-RISK

> The reality is that the chances of a misaligned AI are not small. In fact, in the absence of an effective safety program, that is the only outcome we will get. ... We are looking at an almost guaranteed event with the potential to cause an existential catastrophe. [Yampolskiy (2024)]

In the worst-case scenario, one possibility is for agent *G* to simply decide to **kill all humans**.

Figure 13: A lifeless post-apocalyptic Earth (as imagined by Midjourney)

Given that agent *G* is an AGI Z, and therefore far more intelligent than any human, mere humans would be powerless to stop it. Thus there exists at least one plausible scenario in which a powerful AGI would represent an **existential threat** (**x-risk**) to the continued existence of the human species. Many other scenarios exist where the threat posed would be less than existential but nevertheless **catastrophic** [Hendrycks and Mazeika (2022); Hendrycks, Mazeika, and Woodside (2023)[92]].

2.2.52 THE INTERPLAY BETWEEN INTELLIGENCE AND ALIGNMENT

In Section 2.2.50, we assume that agent *G* is an AGI Z, i.e. the most intelligent system imaginable, and certainly far more intelligent than mere superintelligence (AGI O), or mere humans (AGI L).

If we now **reduce agent *G*'s intelligence** from AGI Z through AGI O to AGI N, AGI L, AGI K, and AGI J — in either the best-case, worst-case, or any intermediate scenario — agent *G*'s **motivations** and **intent** remain the same but its **capabilities** and **effectiveness** are gradually diminished.

As already alluded in Section 2.2.49, **below some minimum threshold of intelligence** (probably ~AGI L, i.e. broadly human-level AGI), agent *G* will lack the cognitive ability to be either maximally-aligned, or maximally-misaligned (i.e. it simply won't understand the necessary concepts, or be able to reason about them as required). **Above that minimum threshold**, however, if agent *G* is (near-)maximally-aligned, then *good* things (resulting from *G*'s actions) will *increase* as intelligence increases — as *G*'s capabilities and effectiveness increase — and *bad* things (resulting from *G*'s actions) will *decrease*. Conversely, if agent *G* is (near-)maximally-misaligned, then *bad* things (resulting from *G*'s actions) will *increase* as intelligence increases — as *G*'s capabilities and effectiveness increase — and *good* things (resulting from *G*'s actions) will *decrease*.

Thus, **as long as agent *G* is (near-)maximally-aligned, the smarter it can be made, the better**.

2.2.53 WELL-FOUNDED AGI

Imagine that we are successful in implementing the best-case (i.e. maximally-aligned) scenario described in Section 2.2.50, with either an AGI O or an AGI Z (either would be totally awesome!)

Given the stakes (if we get it right then we will have created a near-utopia for all mankind for all eternity (Figure 11), but if we get it wrong then we will have created a near-dystopia for all mankind for all eternity (Figure 12)), **we need to be able to convince (i) ourselves, and (ii) everyone else of the correctness of our design and the unimpeachable trustworthiness of its implementation**[93].

It is not enough to simply say "well, we tried it a bunch of times and it seemed to sort of work OK most of the time". Instead, any such system must necessarily be **well-founded** [Russell (2023)]:

- semantically well-defined, individually checkable components
- rigorous theory of composition for complex agent architectures
- built on formally verified [hardware and] software stacks for single and multiple agents.

This requirement will disqualify any implementation technology that cannot satisfy it.

[92] "Rapid advancements in [AI] have sparked growing concerns among experts, policymakers, and world leaders regarding the potential for increasingly advanced AI systems to pose catastrophic risks. ... This paper provides an overview of the main sources of catastrophic AI risks ...: malicious use, in which individuals or groups intentionally use AIs to cause harm; AI race, in which competitive environments compel actors to deploy unsafe AIs or cede control to AIs; organizational risks, highlighting how human factors and complex systems can increase the chances of catastrophic accidents; and rogue AIs, describing the inherent difficulty in controlling agents far more intelligent than humans."

[93] in other words, any such system must necessarily be not just *maximally-aligned*, but **provably maximally-aligned**

2.3 Consciousness

NOTE — This section is incomplete.

2.4 The specific problem that we seek to address

NOTE — This section is incomplete.

2.5 Our proposed approach to the problem that we have identified

NOTE — This section is incomplete.

3. Cognitive Architecture

> It is likely easier for a scientist to explain quantum physics to a mentally challenged deaf and mute four-year-old raised by wolves than for a superintelligence to explain some of its decisions to the smartest human. [Yampolskiy (2024)]

NOTE — This section is incomplete. We provide this snippet as an indication of what is to come!

What should the system's *final goal* be...? We propose **Turner's Three Laws**[94] (T3L):

> "Assume for the purposes of this final goal (expressed in 2024 English) that the physical universe exists; given this assumption, perform Directives D(1) and D(2) (simultaneously, repeatedly, continuously, in perpetuity, and to the best of your ability), subject to Qualification Q(a), with the overall objective of behaving in a manner that is maximally-aligned in perpetuity with the fairly-aggregated idealised preferences of all human beings (living or future), where D(1), D(2), and Q(a) [which are to be interpreted relative to your temporal frame of reference] are as follows:
>
> D(1) for each individual human being $\mathcal{B}$ (living or future), strive to estimate (as accurately as possible) what $\mathcal{B}$'s actual rational well-informed freely-determined preferences $\mathcal{P}$ would be if $\mathcal{B}$ were entirely rational, well-informed, and free from manipulation by any party;
>
> D(2) for each individual human being $\mathcal{B}$ (living or future), strive to maximise the extent to which $\mathcal{B}$'s preferences $\mathcal{P}$ (as estimated pursuant to D(1)) are (and most likely will be) realised;
>
> Q(a) strive to resolve (to the best of your ability, and in a manner that is consistent with a maximally fair aggregation of the individual preferences $\mathcal{P}$ (as estimated pursuant to D(1)) of the living human being population) any conflicts that may arise in respect of D(1) or D(2)."

Note that T3L tacitly assumes that every passive problem-solver underlying the active problem-solver for which T3L is the final goal possesses a minimum level of (a) problem-solving ability (i.e. inventiveness), (b) knowledge and understanding (i.e. information), and (c) available physical resources (e.g. compute); accordingly, a maximally-intrinsic AGI O or Z (which, as already alluded, necessarily possesses (i) an inventor that is better at solving any given problem than any human, (ii) a deep understanding of all human knowledge (grounded by experience)[95], and (iii) sufficient physical resources (e.g. compute) to be able to deliver timely solutions) will amply satisfy this requirement.

We conjecture the following:

- any active problem-solver for which T3L is the final goal will continually re-align itself with human preferences, in perpetuity, as these naturally evolve over future years and centuries
- it's humans (collectively) that are in control of the machine, not the other way around
- no individual human or minority of humans can ever gain control of the machine
- if enough humans genuinely want the machine to switch itself off, then it will do so.

[94] a nod to *Azimov's Three Laws* [Asimov (1942); Asimov (1950)]

[95] including, of course, AI safety, all published AI safety literature, etc; accordingly, (being super-knowledgeable) any such system will not only have a far deeper grasp of AI safety than any human AI safety researcher, but (being super-inventive) it will also be able to formulate far better solutions to any AI-safety-related problem than any human AI safety researcher

4. Construction sequence

NOTE — This section is incomplete.

5. Governance

> For a successful technology, reality must take precedence over public relations, for nature cannot be fooled. Feynman (1986)

NOTE — This section is incomplete.

6. Collaboration

NOTE — This section is incomplete.

7. Conclusion

> It seems a pity, but I do not think that I can write more. For God's sake, look after our people. Scott (1912)

NOTE — This section is incomplete.

8. Acknowledgements

We are deeply indebted to the following informal reviewers, each of whom provided invaluable encouragement and priceless feedback during the course of this paper's interminable development:

Reviewer	Affiliation
Professor Leslie Smith	University of Stirling
Professor Pei Wang	Temple University
Professor Bob Kowalski	Imperial College London

Many thanks to all of you!

References

Adams, D. 1978. The Hitchhiker's Guide to the Galaxy (radio series). BBC Radio 4.

Adler, M. J. 1974. Propædia. In *Encyclopædia Britannica*. Encyclopædia Britannica, Inc., 15th edition.

Allen, J.; Hendler, J.; and Tate, A., eds. 1990. *Readings in Planning*. Morgan Kaufmann.

Arrow, K. J. 1951. *Social Choice and Individual Values*. Yale University Press.

Asimov, I. 1942. Runaround. In *Astounding Science Fiction*. Street & Smith.

Asimov, I. 1950. *I, Robot*. Gnome Press.

Barr, A.; Feigenbaum, E. A.; and Cohen, P. R., eds. 1982. *The Handbook of Artificial Intelligence (Vols 1-3)*. Pitman Books.

Bartha, P. 2022. Analogy and Analogical Reasoning. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition.

Barwise, J. 1977. An introduction to first-order logic. In Barwise, J., ed., *Handbook of Mathematical Logic*. North-Holland.

Benaroya, H., and Han, S. M. 2005. *Probability Models in Engineering and Science*. Taylor & Francis.

Bernardo, J. M., and Smith, A. F. 2000. *Bayesian Theory*. John Wiley & Sons.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

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

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. `https://arxiv.org/pdf/2005.14165.pdf`.

Cole, D. 2023. The Chinese Room Argument. In Zalta, E. N., and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition.

Coons, C., and Weber, M. 2014. *Manipulation: Theory and Practice*. Oxford University Press.

Corke, P. 2011. *Robotics, Vision, and Control: Fundamental Algorithms in MATLAB*. Springer.

Cox, R. 1946. Probability, Frequency, and Reasonable Expectation. *American Journal of Physics* 14(1).

Davies, N. 2009. *Flat Earth News*. Vintage.

Dechter, R. 2003. *Constraint Processing*. Elsevier Science.

Dongarra, J., and Geist, A. 2022. Report on the Oak Ridge National Laboratory's Frontier System. Technical Report ICL-UT-22-05, University of Tennessee.

Douven, I. 2021. Abduction. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Douven, I. 2022. *The Art of Abduction*. MIT Press.

Ernst, G., and Newell, A. 1969. *GPS: A Case Study in Generality and Problem Solving*. Academic Press Inc.

Faul, A. 2020. *A Concise Introduction to Machine Learning*. CRC Press.

Feynman, R. P. 1986. Report of the Presidential Commission on the Space Shuttle Challenger Accident — Appedix F. `https://www.e-education.psu.edu/files/meteo361/file/nasa_report.pdf`.

Flach, P., and Hadjiantonis, A., eds. 2000. *Abduction and Induction: Essays on their Relation and Integration*. Kluwer Academic Publishers.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30(3):411–437.

Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; and Rubin, D. B. 2014. *Bayesian Data Analysis (Third edition)*. CRC Press.

Ghallab, M.; Nau, D.; and Traverso, P. 2016. *Automated Planning and Acting*. Cambridge University Press.

Goertzel, B., and Pennachin, C., eds. 2007. *Artificial General Intelligence*. Springer.

Goertzel, B., and Wang, P., eds. 2007. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*. IOS Press.

Good, I. 1966. Speculations Concerning the First Ultraintelligent Machine. *Advances in Computers* 6:31–88.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press.

Gottfredson, L. S. 1997. Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. *Intelligence* 24(1):13–23.

Halbach, V. 2010. *The Logic Manual*. Oxford University Press.

Halpern, J. Y. 2016. *Actual Causality*. MIT Press.

Halpern, J. Y. 2017. *Reasoning About Uncertainty (Second edition)*. MIT Press.

Handelman, S., ed. 2009. *Thought Manipulation: The Use and Abuse of Psychological Trickery*. ABC-CLIO.

Hansson, S. O., and Grüne-Yanoff, T. 2022. Preferences. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2022 edition.

Harari, Y. N. 2011. *Sapiens: A Brief History of Humankind*. Vintage.

Henderson, L. 2020. The Problem of Induction. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition.

Hendrycks, D., and Mazeika, M. 2022. X-Risk Analysis for AI Research.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks.

Hendrycks, D. 2023. Introduction to AI Safety, Ethics, and Society. `https://www.aisafetybook.com`.

Holland, J. H.; Holyoak, K. J.; Nisbett, R. E.; and Thagard, P. R. 1986. *Induction: Processes of Inference, Learning, and Discovery*. MIT Press.

Jain, S. 1999. *Systems that Learn: An Introduction to Learning Theory*. MIT Press.

Jaynes, E. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Jeffrey, R. 2004. *Subjective Probability: The Real Thing*. Cambridge University Press.

Jeffreys, H. 1939. *Theory of Probability*. Clarendon Press.

Johnson, G. 2016. *Argument & Inference: An Introduction to Inductive Logic*. MIT Press.

Kolmogorov, A. 1933. *Foundations of the Theory of Probability*. Chelsea Publishing Company.

Kowalski, R. 1979. Algorithm = Logic + Control. *Communications of the ACM* 22(1):424–436.

Kowalski, R. 2011a. Artificial Intelligence and Human Thinking (presentation delivered at IJCAI 2011). `https://www.youtube.com/watch?v=k_HGvkfDnT8`.

Kowalski, R. 2011b. *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge University Press.

Lakatos, I. 1976. *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press.

List, C. 2022. Social Choice Theory. In Zalta, E. N., and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Liu, Y.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; Wu, Z.; Zhu, D.; Li, X.; Qiang, N.; Shen, D.; Liu, T.; and Ge, B. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.

MacKay, D. J. 2003. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mendelson, E. 1987. *Mathematical Logic*. Wadsworth & Brooks/Cole, 3rd edition.

Michalski, R. S.; Carbonell, J. G.; and Mitchell, T. M., eds. 1983. *Machine Learning: An Artificial Intelligence Approach*. Tioga Publishing Co.

Mitchell, T. M. 1997. *Machine Learning*. McGraw-Hill.

Monett, D.; Lewis, C. W.; and Thórisson, K. R. 2020. Introduction to the JAGI Special Issue "On Defining Artificial Intelligence" – Commentaries and Author's Response. *Journal of Artificial General Intelligence* 11.

Morris, M. R.; Sohl-dickstein, J.; Fiedel, N.; Warkentin, T.; Dafoe, A.; Faust, A.; Farabet, C.; and Legg, S. 2023. Levels of AGI: Operationalizing Progress on the Path to AGI.

Mortimer, H. 1988. *The Logic of Induction*. Ellis Horwood.

Murphy, K. P. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press.

Newell, A., and Simon, H. A. 1972. *Human Problem Solving*. Echo Point Books and Media.

Newell, A.; Shaw, J.; and Simon, H. A. 1958. Report on a General Problem-Solving Program. Technical Report P-1584, Rand Corporation.

Newell, A. 1990. *Unified Theories of Cognition*. Harvard University Press.

Nilsson, N. J. 2010. *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge University Press.

Noggle, R. 2022. The Ethics of Manipulation. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition.

Oborne, P. 2021. *The Assault on Truth*. Simon & Schuster.

Oppenlaender, J. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference*. ACM.

Pearl, J., and Mackenzie, D. 2019. *The Book of Why: The New Science of Cause and Effect*. Penguin Science.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pólya, G. 1945. *How To Solve It*. Princeton University Press.

Pólya, G. 1954a. *Mathematics and Plausible Reasoning, volume I: Induction and Analogy in Mathematics*. Princeton University Press.

Pólya, G. 1954b. *Mathematics and Plausible Reasoning, volume II: Patterns of Plausible Inference*. Princeton University Press.

Pólya, G. 1962a. *Mathematical Discovery: On Understanding, Learning and Teaching Problem Solving*, volume I. Ishi Press International.

Pólya, G. 1962b. *Mathematical Discovery: On Understanding, Learning and Teaching Problem Solving*, volume II. Ishi Press International.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation.

Rasmussen, C. E., and Williams, C. K. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

Rescorla, M. 2019. The Language of Thought Hypothesis. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition.

Robinson, J. 1965. A Machine-Oriented Logic Based on the Resolution Principle. *Journal of the Association for Computing Machinery* 12 (1):23–41.

Romero, A. 2024. The GPT-4 Generation: Why Are the Best AI Models Equally Intelligent? `https://www.thealgorithmicbridge.com/p/the-gpt-4-generation-why-are-the`.

Rosenthal, J. S. 2006. *A First Look at Rigorous Probability Theory (Second edition)*. World Scientific.

Russell, S., and Norvig, P. 2021. *Artificial Intelligence, A Modern Approach*. Prentice Hall, 4th edition.

Russell, S. 2019. *Human Compatible: AI and the Problem of Control*. Allen Lane.

Russell, S. 2023. How Not To Destroy the World With AI. `https://www.youtube.com/watch?v=ISkAkiAkK7A`.

Schulte, O. 2023. Formal Learning Theory. In Zalta, E. N., and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Scott, R. F. 1912. Sledging diary ('Vol. III'); last entry, 29 March, 1912. `https://www.paulrose.org/news/2012/02/captain-scotts-diaries`.

Searle, J. R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3):417–424.

Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Shoenfield, J. R. 1967. *Mathematical Logic*. Association for Symbolic Logic.

Stone, J. V. 2013. *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis*. Sebtel Presss.

Sutton, R. S., and Barto, A. G. 2018. *Reinforcement Learning*. MIT Press.

Tsang, E. 1993. *Foundations of Constraint Satisfaction*. Academic Press.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 49:433–460.

Venkatesh, S. S. 2013. *The Theory of Probability: Explorations and Applications*. Cambridge University Press.

Walton, D. 2014. *Abductive Reasoning*. University of Alabama Press.

Wang, P., and Goertzel, B., eds. 2012. *Theoretical Foundations of Artificial General Intelligence*. Atlantis Press.

Wang, P. 2013. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. World Scientific.

Wang, P. 2019. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence* 10:1–37.

Wiener, N. 1960. Some Moral and Technical Consequences of Automation (1960). *Science* 131:1355–1358.

Wirth, N. 1976. *Algorithms + Data Structures = Programs*. Prentice-Hall.

Yager, R.; Liu, L.; Dempster, A. P.; and Shafer, G., eds. 2004. *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer.

Yampolskiy, R. V. 2016. *Artificial Superintelligence: A Futuristic Approach*. CRC Press.

Yampolskiy, R. V. 2024. *AI: Unexplainable, Unpredictable, Uncontrollable*. CRC Press.

