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

[unfinished draft]

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

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

Timestamp: 2023-12-07 21:34:18Z

Abstract

The way in which AI — and, in particular, superintelligent AGI — develops over the remainder of this century will most likely determine the subsequent quality of life of all mankind for all eternity. Due to the unique nature of superintelligence, we have one — and only one — chance to get it right. Imminent existential risks notwithstanding, this makes getting superintelligent AGI right the most important problem currently facing mankind. Accordingly, the goal of the BigMother project is to influence the current AGI trajectory (and thus the AGI endgame, and thus the fate of all mankind for all eternity) in order to achieve an endgame that is (as close as possible to) maximally-beneficent (and minimally-maleficent) for all mankind. 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 superintelligent AGI, and then working backwards to build it. To this end, **we seek to design, develop, and deploy a well-founded maximally-aligned maximally-superintelligent alpha AGI (called BigMother/BigMom) that is publicly owned by all mankind, 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. Taken 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)

Overview

This is a long (OK, book-length) paper, whose structure may be summarised as follows:

- In Section 1, we:
  - explain why this paper is in two parts
  - introduce a number of basic concepts and definitions pertaining to AGI[1]
  - describe the specific AGI-related problem that we seek to address
  - describe a two-pronged approach to the problem that we have identified.
- In Section 2, we describe a cognitive architecture for a proposed "Gold Standard" AGI
  - this section includes a tentative solution to the *AI alignment* problem.
- In Section 3, we explore the concept of consciousness in relation to AGI.
- 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:
  - restate the original problem
  - summarize our overall argument
  - discuss the implications.
- In Section 8, we thank our informal reviewers for their contributions.

[1] Artificial General Intelligence

Contents

1. Introduction

This paper describes work conducted by the author from 1985 to 2023, but not previously reported[2].

1.1 It takes a village

Artificial General Intelligence (AGI) [Turing (1950); Goertzel and Pennachin (2007); Goertzel and Wang (2007); Wang and Goertzel (2012); AGI Society (2009 23)] — 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 both 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, two variants of this paper are currently envisaged, as described in Table 1:

Table 1: Variants of this paper

Variant	Accessible?	Description
Part 1	yes	assumes only basic high-school mathematics (plus some determination!)
Part 2	no	expands upon Part 1; contains more advanced mathematics

Although quite long (currently 63 pages), Part 1 is nevertheless designed to be an easy read. Thus, given the complexities of AGI, Part 1 is a good starting point even for more advanced readers.

1.2 Basic concepts and definitions

1.2.1 WHAT IS INTELLIGENCE?

It's not really possible to answer questions about intelligence, Artificial Intelligence (AI), or Artificial General Intelligence (AGI) without first considering what these things mean. Despite prior attempts at providing formal definitions [e.g. Barr, Feigenbaum, and Cohen (1982); Gottfredson (1997); 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.

1.2.2 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

1.2.3 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 highly subjective) understanding of "intelligence") on the basis of its externally observable behaviour[3].

1.2.4 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 somehow *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.

1.2.5 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.

1.2.6 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 entirely 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)

1.2.7 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 *passive* problem-solvers[5] and *active* problem-solvers. A *passive problem-solver* takes as input some kind of *problem statement* or *problem description*, 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*[6].

1.2.8 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[7]
- R tells us how much time, energy, compute, money etc we can use in the search for a solution[8].

A triple $\langle P, Q, R \rangle$ should be interpreted by a passive problem-solver as follows: "using no more than resources R, strive to find some Q-maximal thing x having property P"; e.g.[9]:

- 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] broadly comparable to an *Oracle AI* [see e.g. Armstrong, Sandberg, and Bostrom (2012)]
[6] active problem-solvers (often referred to as *Agentic AI*) will be described more fully in Section 1.2.26
[7] if Q is not specified, then any valid solution is considered to be as good as any other
[8] in particular, in many real-world contexts, only *timely* solutions have any significant utility [see e.g. Newell (1990)]
[9] 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 $\langle P, Q, R \rangle$ will be described using a formal notation

1.2.9 PROBLEM SOLUTIONS

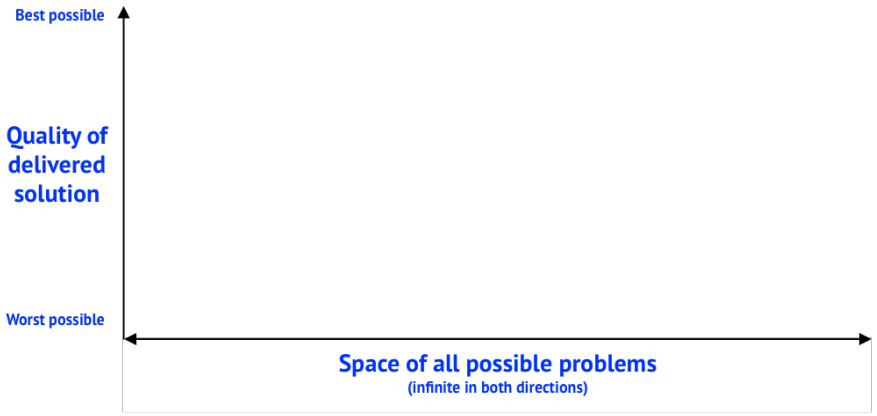
Given an arbitrary problem $\langle P, Q, R \rangle$, a passive problem-solver:

- will strive to find the best valid solution that it can, with no guarantees beyond that
- may deliver any of the following[10]:
  - 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[11]. Note that some passive problem-solvers will deliver better (higher-quality) solutions than others.

1.2.10 SOLUTION QUALITY

If we set the finite time limit within which problem solutions must be found to some arbitrary fixed value, but otherwise allow a problem-solver unlimited 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 delivered solutions according to ordering Q (on the y-axis). We can then plot different passive problem-solvers on the same 2D graph, for comparison[12]:



(For the avoidance of doubt, the following graphs (depicted in Sections 1.2.11 to 1.2.16) pertain specifically to *passive problem-solvers* (as defined in Section 1.2.7), assuming a fixed time limit.)

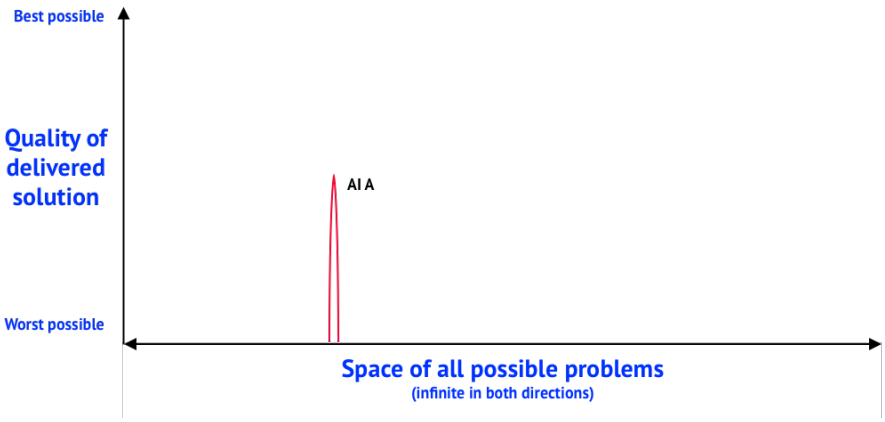
[10] for our present purposes, we make the simplifying assumption that any passive problem-solver is *deterministic* (given the same $\langle P, Q, R \rangle$, it always delivers the same result); in reality, many passive problem-solvers will be *nondeterministic*

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

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

1.2.11 NARROW AI

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

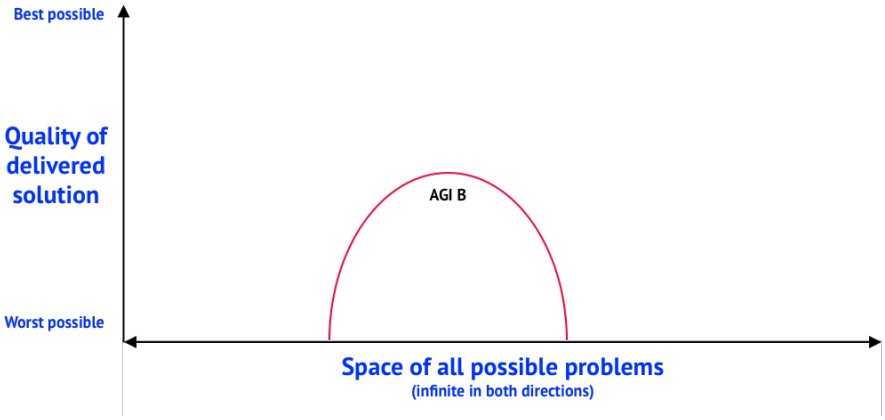


Here, AI A is a narrow AI (according to our definition).

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.

1.2.12 ARTIFICIAL GENERAL INTELLIGENCE (AGI)

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

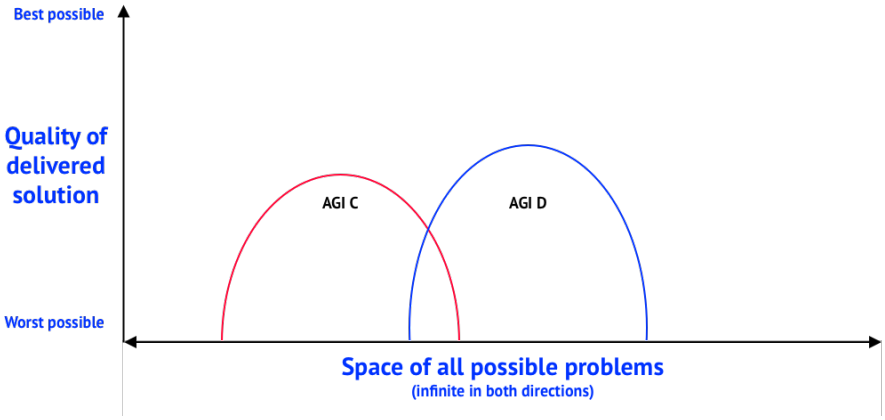


Here, AGI B is an AGI (according to our definition).

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.

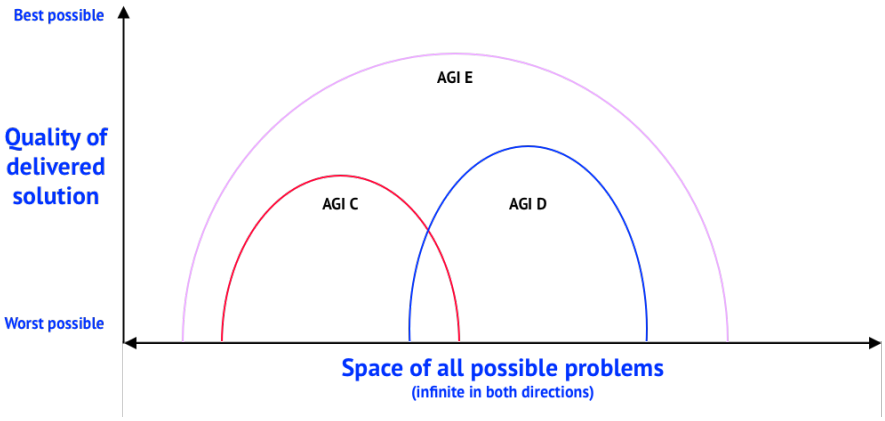
Some AGIs might be particularly adept in some problem areas, but not in others:



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

1.2.13 THE ALPHA AGI

It's also possible for one AGI (the "alpha") to outperform all its peers across all possible problems:

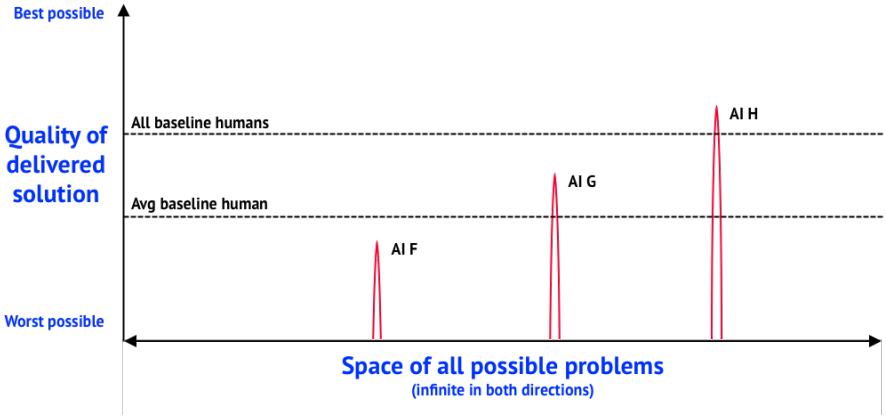


Here, AGI E outperforms both AGI C and AGI D at all things.

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

1.2.14 SUPERINTELLIGENT NARROW AI

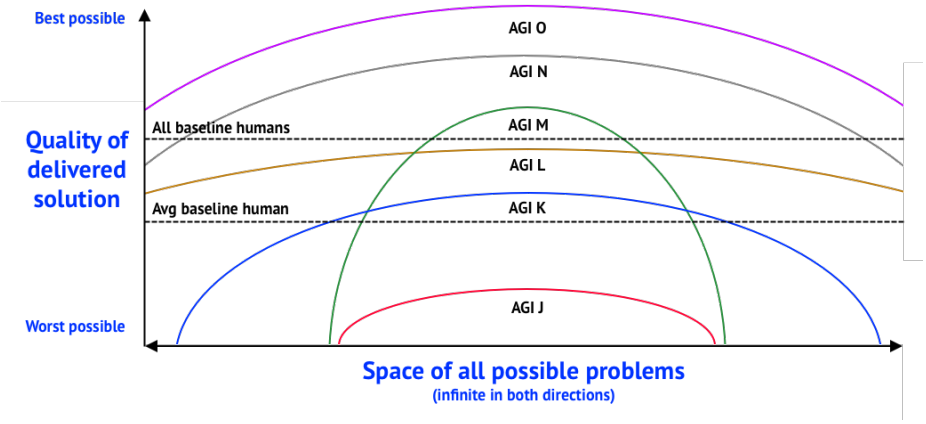
Things get interesting when we compare AI performance against *baseline human* performance[13]:



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

1.2.15 SUPERINTELLIGENT AGI

We can also compare AGI performance against baseline human performance:

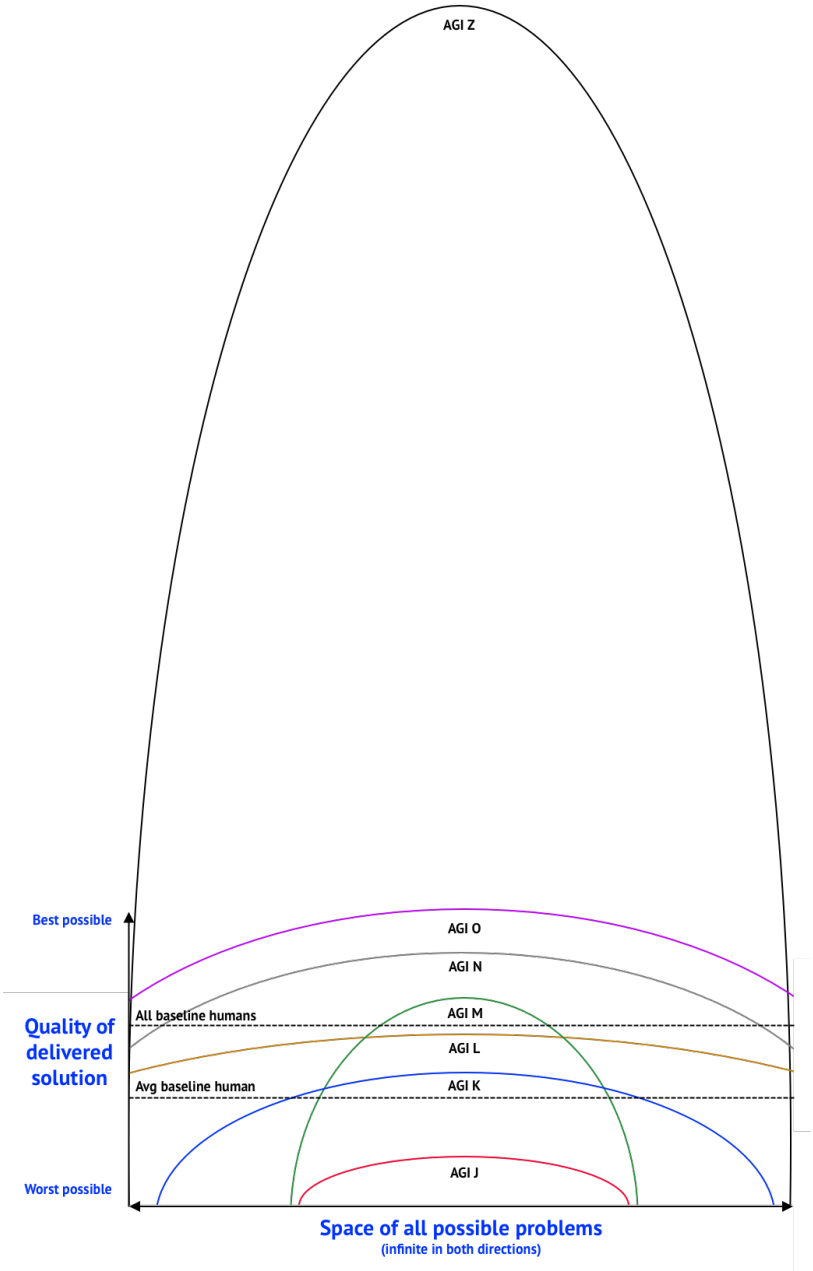


There are six primary possibilities. The average baseline human outperforms AGI J at everything, AGI K outperforms the average baseline human at some things, AGI L outperforms the average baseline human at everything[14], AGI M outperforms all baseline humans at some things and the average baseline human at more things, AGI N outperforms all baseline humans at some things and the average baseline human at everything, and AGI O outperforms all baseline humans at all things. In other words, AGI O (here, the alpha AGI) is *superintelligent* across all possible problems[15].

[13] by "baseline human" we mean a human unassisted by any other system that would itself qualify as an AI as above
[14] broadly speaking, AGI L seems to be what most AI researchers mean by "human-level AGI" [e.g. Morris et al. (2023)]
[15] a sufficiently performant AGI M or AGI N might be judged to be *near-superintelligent*

1.2.16 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 (AGI O) is a lower bound.

1.2.17 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

1.2.18 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.

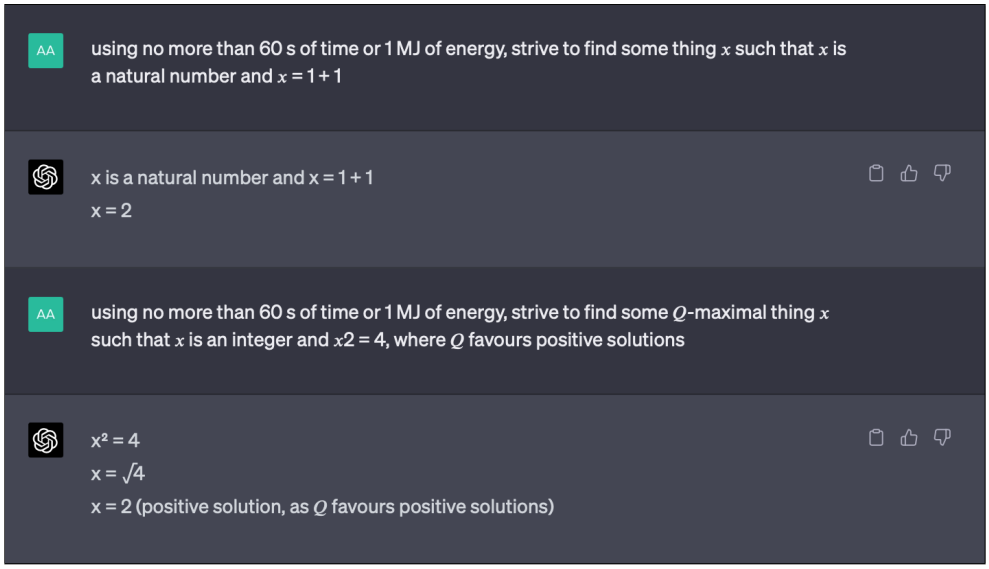
Is it superintelligent? No — we would judge DALL·E to be an AI G, far short of AI H.

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

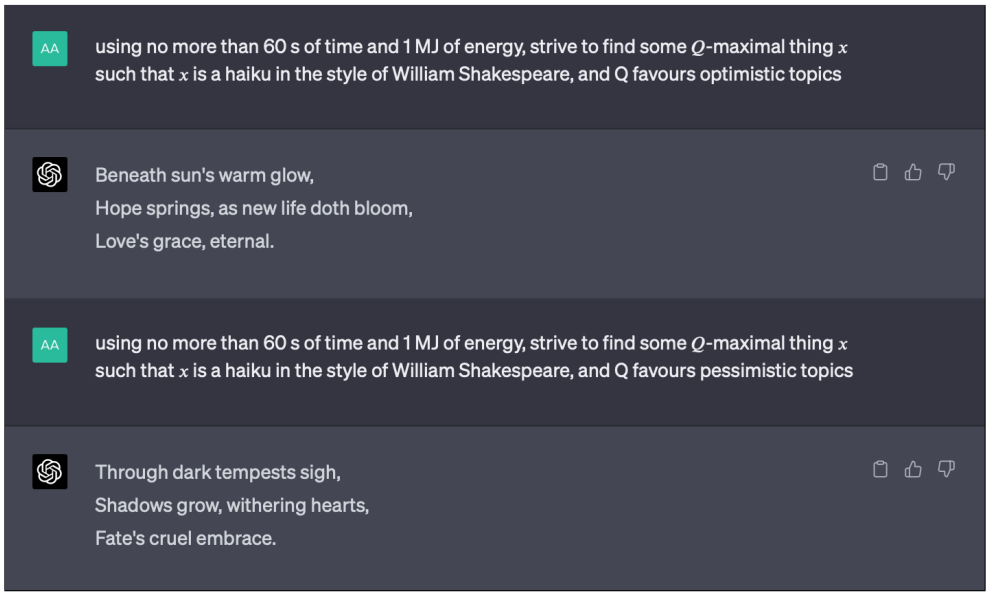
1.2.19 EXAMPLE — CHATGPT

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

Example — arithmetic:

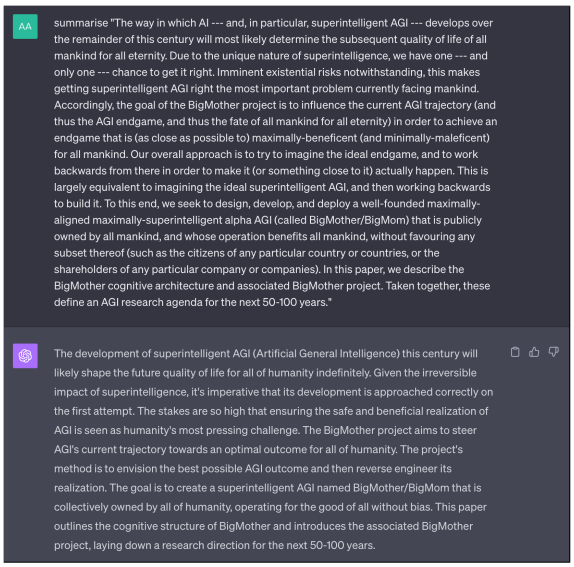


Example — haikus[23]:



[21] specifically, the mid-2023 version based on the GPT-4 LLM, without any external embellishments such as plug-ins
[22] note that Google's Bard chatbot (powered by the Gemini LLM released in December 2023) is broadly comparable
[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 problem 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.

1.2.20 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 not 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!

[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 two AI systems, 60 years apart, both stalling at AGI K, suggests that AGI K is a difficult barrier to overcome!

[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)

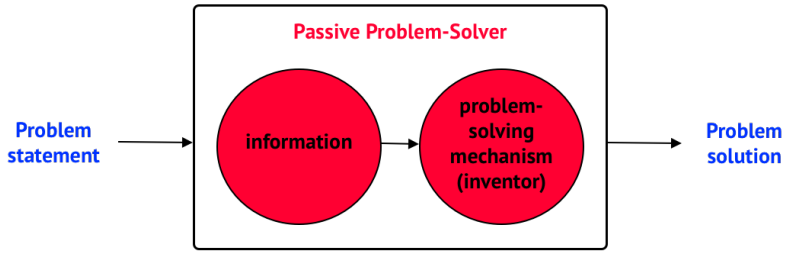
1.2.21 THE ROLE OF INFORMATION IN PROBLEM-SOLVING

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* that drives the inventor towards solutions:



The information that drives the inventor towards solutions must be:

- *acquired* (externally) somehow[29]
- *represented* (internally) somehow[30]
- *processed* (by the inventor) somehow[31].

1.2.22 MAXIMALLY-INTRINSIC PASSIVE PROBLEM-SOLVERS

Earlier, in Section 1.2.20, 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 whatever created the massive lookup table, e.g. a human AI designer[32].

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

[29] passive problem-solvers don't do any *information acquisition* themselves (it's all done for them by some other party)

[30] for a wannabe superintelligent AGI, the *information representation mechanism* (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

[31] for example, via some arbitrarily complex combination of *induction*, *deduction*, and *abduction* (see Section 1.2.25)

[32] for the time being, we shall mostly restrict our discussion to the case where a *human being* is designing an *AI system*

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

1. consider every possible problem statement (i.e. every possible problem instance)[33]
2. attempt to solve each problem instance themselves (e.g. by hand-executing a suitable algorithm)
3. add the results to the lookup table.

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]:

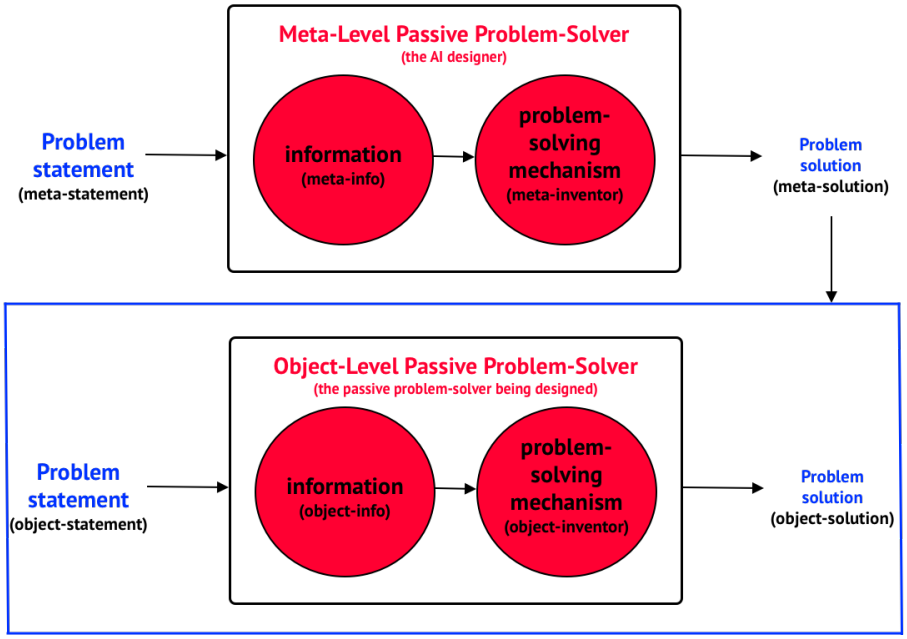


Figure 1: A meta-level passive problem-solver designing an object-level passive problem-solver

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

meta-statement	the meta-level problem statement (potentially 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
meta-solution	the meta-level problem solution (wrt the specified meta-statement)
object-statement	the object-level problem statement
object-info	other information that the object-level problem-solver has in its possession
object-inventor	the object-level problem-solving mechanism
object-solution	the object-level problem solution (wrt the specified object-statement)

[33] even if the number of possible problem statements is infinite, we imagine that this is somehow magically possible
[34] although our primary focus is on AI, 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

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** 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.

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[36].

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
    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.

[36] in other words, a maximally-intrinsic (maximally-intrinsically intelligent) 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

1.2.23 MAXIMALLY-INTRINSIC SUPERINTELLIGENT AGI

Any maximally-intrinsic superintelligent passive problem-solver (AGI O) must necessarily possess:

1. *a deep understanding of all human knowledge* (grounded by experience)[37]
2. an inventor that is *better at solving any given problem than any human*
3. sufficient physical resources (e.g. compute) to be able to deliver *timely* problem solutions.

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

$$\textit{superintelligent} = \textit{super-knowledgable} + \textit{super-inventive} + \textit{super-resourced}^{38}.$$

1.2.24 THE PATH TO SUPERINTELLIGENT AGI

More generally, *intelligence* has three scalable dimensions: *knowledge* + *inventiveness* + *resources*[39].

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 "*knowledge* + *inventiveness* + *resources*" combination of $\mathcal{H}$'s meta-information + $\mathcal{H}$'s meta-inventor + "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-information + $\mathcal{M}+$'s object-inventor + 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[40].

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. Consequently, 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.

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

1.2.25 INDUCTION, DEDUCTION, AND ABDUCTION (IDA)

It seems to be the case that — at the highest levels of abstraction (i.e. at the level of *mind* rather than *brain*) — critical thought and problem-solving *within humans* involves three modes of reasoning:

- **induction**[41, 42] — 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**[43] — 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**[44] — 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[45].

Thus induction, deduction, and abduction (IDA)[46] may be viewed as *cognitive primitives* on top of which higher-level reasoning such as generic problem-solving may be constructed. Specific formulations of IDA may be either *strong* (well-founded) or *weak* (prone to some kind of error), e.g.:

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

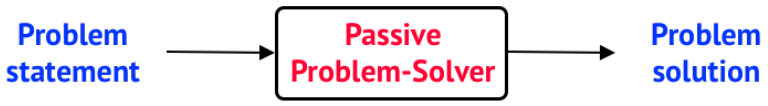
We conjecture the following[47]:

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) will require *strong* IDA
5. maximally-superintelligent AGI (AGI Z) will require *strong* IDA.

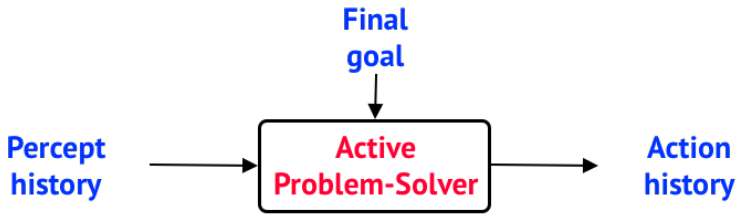
[41] Holland et al. (1986); Mortimer (1988); Johnson (2016); Henderson (2020); Bartha (2022)
[42] note that *induction* and *analogy* are closely related, a subject to which we shall return in Section ???
[43] Shoenfield (1967); Barwise (1977); Mendelson (1987); Halbach (2010)
[44] Walton (2014); Douven (2021); Douven (2022)
[45] these are *plausible hypotheses* only; elevation to *unqualified belief* (e.g. Socrates *is* a cat) requires additional evidence
[46] however implemented, e.g. either explicitly, or as emergent properties of some massively parallel holistic system
[47] it will likely take many decades of research for these conjectures to be tested, but nevertheless these are our suspicions

1.2.26 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 2.

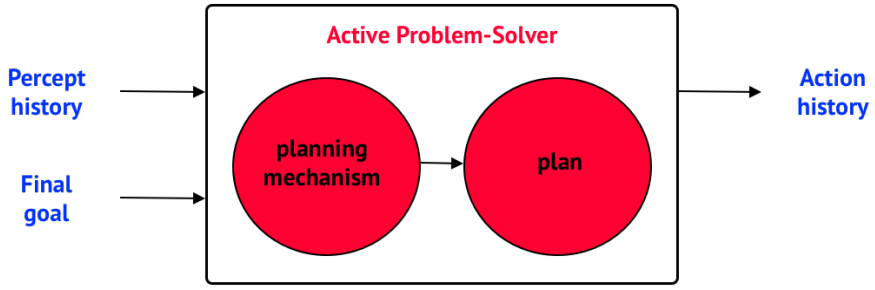
Note that, whereas a *passive* problem-solver must attempt to solve a specified problem $\langle P, Q, R \rangle$ using just the information that it already has and the physical resources R specified by the problem statement, an *active* problem-solver may acquire *additional information* and *additional resources* through interaction with the physical universe via its sensors and effectors.

1.2.27 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[48].

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 state of the (assumed) physical universe (as indicated by the current percept history), formulate a plan which, when executed, will generate a sequence of actions (to be appended to the action history) whose most likely net effect is to make 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 computer programmer, or to a human AI designer (meta-level passive problem-solver).

Accordingly, following Section 1.2.22, 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)[49].

As before, there are many ways in which the AI designer may choose to solve this problem. For example, it might be possible, in some cases, to reformulate the specified final goal as a finite set of *production rules*[50] to be blindly followed in response to incoming percepts, thereby generating actions "whose most likely net effect is to make progress towards (and ideally achieve) the final goal".

In this configuration, 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 1.2.24, 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.

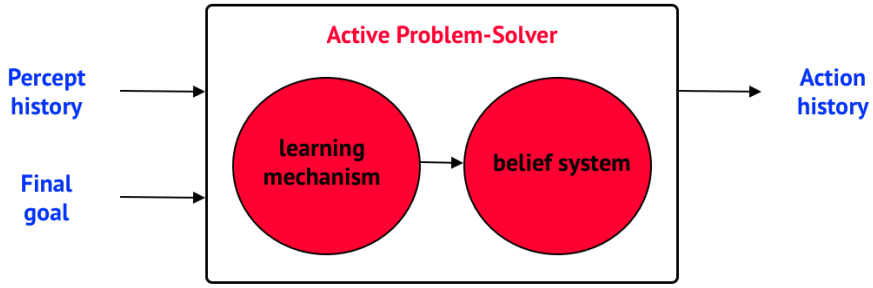
[48] in the simplest possible case, a planning mechanism might generate a plan which generates a single action at a time

[49] 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)

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

1.2.28 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, formulate a belief system (set of beliefs) capturing (as accurately as possible) the structure of the (assumed) physical universe, including all of its nuanced complexity". A belief system, synthesised in this way, is effectively an internal model of the universe[51]. Due to the central 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[52].

Again following Section 1.2.22, the body of the learning loop is effectively an object-level passive problem-solver to be designed by a meta-level passive problem-solver (human AI designer).

As before, there are many ways in which the AI designer may 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.

[51] encompassing, 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

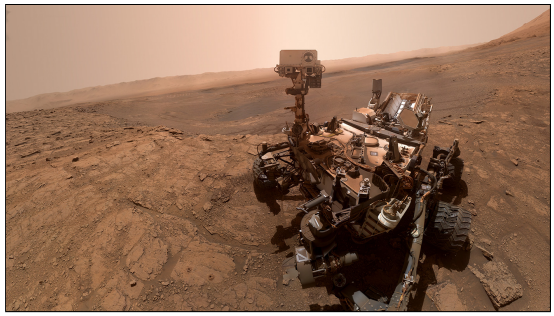
[52] this quality of *adaptability* is often cited in the literature as a core principle of intelligence [see e.g. Wang (2013)]

1.2.29 AUTONOMOUS ROBOTS

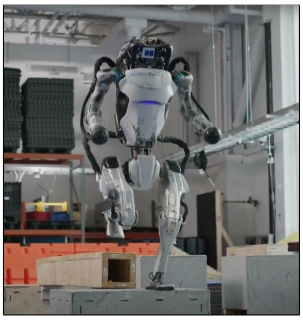
Being a physical system (ultimately comprising computer hardware etc), an active problem-solver must necessarily possess some kind of physical "body", including its physical sensors and effectors.

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 2a) is capable of surface locomotion.



(a) *Curiosity* takes a selfie



(b) *Atlas* goes for a run

Figure 2: 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[53] and effectors[54] (e.g. a Boston Dynamics *Atlas* — Figure 2b)
2. a maximally-intrinsic active problem-solver[55] (Section 1.2.26)
   (a) incorporating
      (i) a continuous planning mechanism (Section 1.2.27)
         — itself incorporating a superintelligent passive problem-solver (AGI O)
      (ii) a continuous learning mechanism (Section 1.2.28)
         — also incorporating a superintelligent passive problem-solver (AGI O)
   (b) that is *maximally-aligned with human beings in perpetuity*.

We shall explore the concept of *maximal alignment* over Sections 1.2.30 to 1.2.39.

[53] 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

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

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

1.2.30 LIVENESS AND SAFETY

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

- autonomous robots
- organisations.

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

1. **good (i.e. desirable) things happen**[56] — usually referred to as *liveness*
2. **bad (i.e. undesirable) things don't** — usually referred to as *safety*.

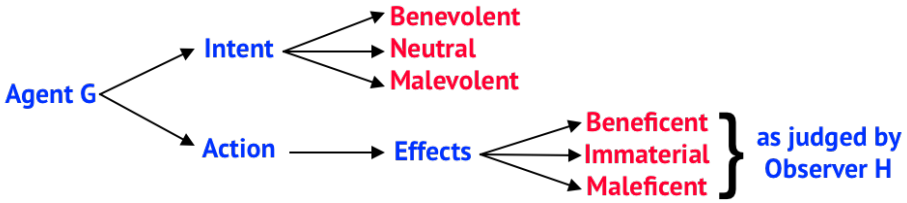
Things are a little more complicated than might appear at first glance because:

- the concepts of *desirable* and *undesirable* are *subjective* (i.e. relative to some observer)
- *intent* and *effect* are two different things.

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

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

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



Furthermore, different observers H_1 and H_2 might judge the effects of G's actions differently.

[56] as a result of G's actions

1.2.31 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, more generally, some *population of observers* $\mathcal{H}$[57]. An agent G is *aligned with human values* if G's behaviour is consistent with *human values* $\mathcal{V}$[58].

1.2.32 HUMAN VALUES

If G is a maximally-intrinsic agent (such as an autonomous robot, or an organisation) which we (as G's designers) desire to be aligned with human values, 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 might be achieved:

(1) we (as G's designers) work out what human values $\mathcal{V}$ to use, and "code them" into G[59], 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 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 $\mathcal{V}$. One advantage of this approach is that agent G *continually re-aligns itself in perpetuity* as human values $\mathcal{V}$ evolve over time.

Figure 3: Human values (as imagined by Midjourney)

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

1.2.33 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 effects* $E_1, ..., E_n$ with corresponding *probabilities* $P_{\mathcal{A},1}, ..., P_{\mathcal{A},n}$[60].

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 (where any strict order is irreflexive[61], asymmetric[62], and transitive[63]) such that $S_x < S_y$ if and only if individual human observer H judges state S_x to be *less desirable* than state S_y. Such an ordering[64] is said to express H's *individual preferences* [Arrow (1951); Hansson and Grüne-Yanoff (2022)]. If the initial state is S_I then effect E corresponding to new state S_E is beneficent[65] if $S_I < S_E$, and maleficent[65] if $S_E < S_I$. If states are partially (rather than totally) ordered then some effects may be immaterial[65], i.e. neither beneficent nor maleficent:

Beneficent effect E:

Immaterial effect E:

Maleficent effect E:

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

1.2.34 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. Analogously to ordinal preferences, if the initial state is S_I then 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 if $U(S_E) = U(S_I)$[66].

The *expected utility* (according to U) of an action sequence $\mathcal{A}$ having possible effects $E_1, ..., E_n$ (corresponding to new states $S_1, ..., S_n$) with corresponding probabilities $P_{\mathcal{A},1}, ..., P_{\mathcal{A},n}$ may be calculated by multiplying, for each possible effect E_i, the utility (according to U) of the corresponding new state S_i by the probability $P_{\mathcal{A},i}$ that the effect in question will occur, and then summing the results:

$$\textit{expected utility} = [U(S_1) \times P_{\mathcal{A},1}] + [U(S_2) \times P_{\mathcal{A},2}] + [U(S_3) \times P_{\mathcal{A},3}] + ... + [U(S_n) \times P_{\mathcal{A},n}].$$

[60] in other words, $P_{\mathcal{A},i}$ is the probability (between 0 and 1) that effect E_i will occur as a result of action sequence $\mathcal{A}$

[61] $a \not< a$

[62] if $a < b$ then $b \not< a$

[63] if $a < b$ and $b < c$ then $a < c$

[64] which may be *total* (such that a preference always exists between any two distinct states), but may also be *partial*

[65] as judged by H

[66] note how the calculations are such that *impact* (either beneficent, maleficent, or immaterial) is always ≥ 0

1.2.35 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.
- Similarly, 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.
- Accordingly, H's *stated* preferences are not necessarily the same as H's *actual* preferences[67].
- It's possible that H is less than entirely well-informed[68] 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) 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 other party.

[67] just ask *The Sorcerer's Apprentice* [Wiener (1960)] or *King Midas* [Russell (2019)]
[68] here, a *well-informed* person does not simply have *access* to the relevant information, but also fully *understands* it

1.2.36 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 $\mathcal{H}$[69] 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 other 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[70]. In other words, there is no unambiguously "perfectly fair" method of doing so[71] — 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)[72].

1.2.37 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, e.g. the next M actions
- G attempts to predict, for every action sequence $\mathcal{A}$ of length M:
  - the possible effects (of action sequence $\mathcal{A}$) $E_1, ..., E_n$
  - their corresponding probabilities $P_{\mathcal{A},1}, ..., P_{\mathcal{A},n}$
- G then performs whichever action sequence yields 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 $\mathcal{H}$.

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

[70] 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 calibration* requires additional information

[71] 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 required aggregated utility function $\hat{U}$

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

1.2.38 PERFECT ALIGNMENT

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

1. the total impact (according to U) of the *beneficent* effects of G's actions are *maximised*, and
2. the total impact (according to U) of the *maleficent* effects of G's actions are *minimised*[73].

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 effects of its actions will never be entirely accurate.

1.2.39 MAXIMAL ALIGNMENT

We will say that agent G is *maximally-aligned* with population $\mathcal{H}$ (such as the population of all humans) 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)[74].

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 natural language (e.g. English), humans, human society, human psychology, human drives and emotions, human values, social choice theory, rationality, self-determination, mathematics, physics, and 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 be able to reason about all of the above, G will need to be able to apply robust critical thought and problem-solving (strong 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[75]; 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.

[73] analogously, *perfect misalignment* = total beneficent impact is minimised and total maleficent impact is maximised
[74] analogously, *maximal misalignment* = the distance from perfect misalignment is minimised (consistent with $-U$)
[75] we suspect easily > 10 orders of magnitude more compute per USD (and ideally per W) than is available in 2023

1.2.40 ALIGNMENT IS EVERYTHING!

Compare the following best- and worst-case scenarios:

- **best-case scenario** — imagine that agent G is an AGI Z (maximally-super-knowledgeable + maximally-super-inventive + maximally-super-resourced, i.e. the most intelligent superintelligent AGI that it's possible to build)[76], 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 4: 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.

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

Figure 5: 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**.

1.2.41 X-RISK

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

Figure 6: 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 scenario in which a powerful AGI would represent an existential threat to the continued existence of the human species (a.k.a. *x-risk*). Many intermediate scenarios, in between the best and worst cases, may also represent an x-risk via a variety of mechanisms [Hendrycks and Mazeika (2022); Hendrycks, Mazeika, and Woodside (2023)[77]].

1.2.42 THE INTERPLAY BETWEEN INTELLIGENCE AND ALIGNMENT

In Section 1.2.40, 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 gradually *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 1.2.39, below some minimum threshold of intelligence (probably ~AGI L), 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, if agent *G* is maximally-aligned (or 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 maximally-misaligned (or 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[78].

1.2.43 WELL-FOUNDED AGI

Imagine that we are successful in implementing the best-case (i.e. maximally-aligned) scenario described in Section 1.2.40, with either an AGI Z or an AGI O (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 4), but if we get it wrong then we will have potentially created a near-dystopia for all mankind for all eternity (Figure 5)), 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.

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.

[77] "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."

[78] in other words, if you're going to build a maximally-aligned AGI, you might as well make it maximally-superintelligent!

1.3 ********** HERE BE DRAGONS **********

A note to the reader:

- everything up to this point is more-or-less locked down
- everything from this point on is still very much up in the air and subject to constant change
- extreme caution is advised if proceeding any further!

1.4 Motivation

Now that we have a basic vocabulary for AI and AGI, we can describe the specific problem that we seek to address. As a first step, it will be instructive to review mankind's evolutionary, migratory, and cultural history [Aydon (2007); Harari (2011); Sapolsky (2018); Cassidy (2021); Roberts (2023)].

1.4.1 LIFE ON EARTH

A recent paper [Wong et al. (2023)] asserts that phenomena such as stellar evolution (the creation of stars and planets), nucleosynthesis (the creation of new atomic nuclei), mineral evolution (the creation of complex molecules), and biological evolution (the creation of life, consciousness, and intelligence) are all *emergent properties* of the causality-driven evolution of the physical universe, and all explained by a single underlying principle, namely the "law of increasing functional information".

Soon after the Sun and Earth formed ~4.5 billion years ago, Earth's atmosphere comprised primarily ammonia, carbon dioxide, hydrogen, hydrochloric acid, hydrogen sulphide, methane, nitrogen, and water vapour. Early Earth was extremely hot, with temperatures in excess of 200°C; accordingly, there was no surface water. ~3.8 billion years ago, the Earth cooled below 100°C, allowing the water vapour in the atmosphere to condense into rain, thereby forming Earth's oceans.

Prokaryotes — single-celled DNA-based organisms lacking cell nuclei — emerged roughly 300,000 years later, i.e. ~3.5 billion years ago. As these organisms multiplied over the next billion years, some of them (cyanobacteria) produced an oxygen-rich atmosphere through photosynthesis (the Great Oxygenation Event), triggering the mass extinction of many organisms for whom oxygen was toxic, as well as a series of ice ages (the Huronian glaciation) from ~2.5 to ~2.2 billion years ago.

1.4.2 MANKIND'S EVOLUTIONARY, MIGRATORY, AND CULTURAL HISTORY

Life, nevertheless, continued[79].

~2.1 billion years ago, we (that is, mankind's very early evolutionary ancestors) were *eukaryotes* — DNA-based organisms comprising one or more cells, each having a membrane-bound nucleus[80].

We somehow survived the Cryogenian ice age which lasted from ~720 to ~630 million years ago.

~610 million years ago, we became[81] *animals* — multi-cellular eukaryotes that (a) consume organic material[82], (b) breathe oxygen, (c) are able to move independently, and (d) reproduce sexually.

~505 million years ago, we became *vertebrates* — animals having (a) a rigid axial vertebral column (spine), (b) a brain and spinal chord, and (c) a closed circulatory system (blood, heart, etc).

We somehow survived the Andean-Sahara ice age from ~460 to ~420 million years ago.

~445 million years ago, we somehow survived the Ordovician–Silurian extinction events that killed around 85% of all species, most likely the result of global cooling and reduced sea levels.

Up until this point in our evolutionary history, we had lived in Earth's oceans. From ~418 to ~395 million years ago, we evolved from fish into land-based *tetrapods* — vertebrates having four limbs.

We somehow survived the late Paleozoic ice age from ~360 to ~255 million years ago.

~252 million years ago, we somehow survived the Permian–Triassic extinction event (the Great Dying) that killed around 90% of all species, most likely the result of massive volcanic eruptions that released sulfur dioxide and carbon dioxide, elevating global temperatures, and acidifying the oceans.

[79] the remainder of this section is a greatly simplified approximation, but nevertheless sufficiently accurate for our purposes
[80] eukaryotes are believed to have evolved from prokaryotes via a process known as *symbiogenesis*
[81] i.e. "evolved (via biological evolution as per Wong et al. (2023)) into"
[82] including, in some instances, other animals!

~220 million years ago, we became *mammals* — warm-blooded homeothermic tetrapods having (a) mammary glands (producing milk for newborn offspring), (b) a neocortex (the outer six layers of the cerebral cortex), (c) fur or hair, and (d) three middle ear ossicles (malleus, incus, and stapes).

~201 million years ago, we somehow survived the Triassic–Jurassic extinction event that killed around 75% of all species, most likely the result of volcanic eruptions as per the Great Dying.

~66 million years ago, we somehow survived the Cretaceous–Paleogene extinction event (a.k.a. the K–T extinction event) that killed around 75% of all plant and animal species, including all non-avian dinosaurs, as a result of the Chicxulub asteroid impact off the Yucatán Peninsula in Mexico.

~65 million years ago, we became *primates* — mammals having (a) large brains relative to body mass, (b) improved visual acuity (forward-facing eyes, stereoscopic vision, excellent depth perception, and (usually) colour vision), (c) a high degree of shoulder mobility, (d) grasping hands with five digits and sensitive tactile pads, (e) claws that have been modified into flattened nails, (f) typically only one offspring per pregnancy, and (g) a trend toward holding the body upright. Most primates spend their lives in complex, tightly woven societies and communicate with each other via smells, touching, body language, visual messages, and sounds. Early primates lived in tropical rainforests.

~25 million years ago, we became *apes* — primates lacking tails but having (a) opposable thumbs, and (b) short lumbar vertebrae able to bear the weight of the upper body while sitting or standing.

By ~20 million years ago, Earth supported a variety of land habitats, including both tropical forests and open grasslands (much like today's savannahs), favouring species able to adapt to both.

~18 million years ago, we became *hominids* (great apes). Even primitive (non-human) hominids exhibit skills and intelligence well beyond the capabilities of their earlier ape ancestors. For example: (a) they make complex nests for sleeping at night, (b) they practice strategic planning in their social lives, (c) gorillas have been known to use stones as both hammer and anvil to crack nuts, (d) orangutans are able to prise fruit open with a stick, and frequently use large leaves to protect themselves from the rain, and (e) chimpanzees use grass stems or sticks to extract termites from their mounds, and often make "sponges" from leaves in order to extract drinking water from holes in trees. Chimpanzees also hunt cooperatively, and wage war against neighbouring chimpanzee communities [Goodall (2010)].

~9 million years ago, we diverged from chimpanzees (our closest hominid relative), with whom we share ~99% of DNA (however, the fossil record from that period is very sparse, so a lot is unknown!)

~3.9 million years ago, we (most likely) became *Australopithecus afarensis*[83], with a mixture of ape-like and human-like features (for example, walking on two legs), somewhere in East Africa.

Figure 7: The Laetoli footprints were most likely made by Australopithecus afarensis (A. afarensis)

Recent finds suggest that A. afarensis may have made and used stone tools ~3.4 million years ago.

[83] or possibly some other variant of Australopithecus

We somehow survived the Quaternary glaciation (an alternating series of glacial and interglacial periods) that began ~2.58 million years ago (with the most recent cycle ending ~11,500 years ago).

~2.5 million years ago, we became *Homo habilis*[84] (H. habilis) — **the first humans** — with an average brain size of 610 cm^3 (approximately 35% larger than that of A. Afarensis) and correspondingly enhanced cognitive abilities. We lived in Africa on the sub-Saharan savannah in polygynous groups of ~80 (with each male having multiple mates); competition for mates frequently led to intense male-male conflict. We survived on a diet of fruit and (mostly scavenged) meat, using early Oldowan tools[85] for butchering and skinning carcasses, as well as for crushing bones (to extract the marrow).

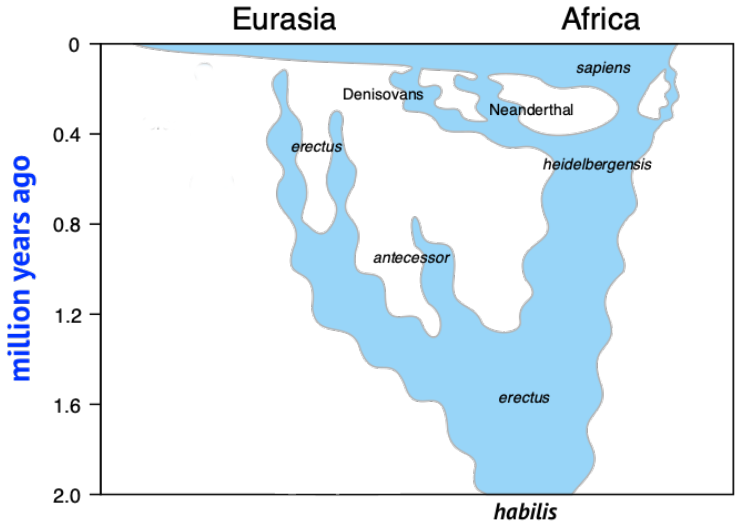


Figure 8: Timeline of post-habilis evolution and migration (derived from Stringer (2012))

~2 million years ago, we (H. habilis) became *Homo erectus*[86] (H. erectus), with an average brain size of 950 cm^3. Although we lacked the vocal apparatus of modern humans, we likely communicated via some spoken proto-language, thereby facilitating intra-group planning and coordination. We dispersed throughout Africa, and successfully migrated to Western Asia, then East Asia and Indonesia, mostly by foot, but possibly also by sea. We lived in largely monogamous groups, thereby reducing the incidence of male-male conflict over mates, as well as facilitating both parental care (sharing the load between two parents) and social cohesion within groups. Being highly-coordinated hunter-gatherers, with evidence of both teamwork and specialisation, we became the apex predator, existing on a diet of mostly meat (mostly hunted rather than scavenged) supplemented by fruits, vegetables, seeds, nuts, and tubers. We made and used Acheulean tools[87] (e.g. hand axes) which require skill, imagination, and planning in their manufacture. There is evidence that we arranged rock piles in order to form simple windbreaks — the earliest evidence of architecture. We may have used fire (opportunistically, exploiting natural wildfires rather than starting our own fires) for cooking, as well as for artificial light, leading to extended waking hours as well as increased intra-group socialisation. There is also evidence that, within our own groups, we cared for the sick, injured, and possibly even the elderly.

[84] "handy man"
[85] also called Mode 1 tools — *** explanation
[86] "upright man"
[87] also called Mode 2 tools — *** explanation

~1.2 million years ago, we (H. erectus) gave rise to *Homo antecessor*[88] (H. antecessor) in some unknown location. H. antecessor eventually reached Northern Spain, and possibly much of Western Europe. Although H. antecessor hunted in organised groups, there is no evidence of the use of fire. There is, however, evidence of simple tool use, conflict with neighboring tribes, and cannibalism.

~800,000 years ago, the last surviving H. antecessor went extinct somewhere in Europe.

~700,000 years ago, we (that is, *African* H. erectus[89], specifically) became *Homo heidelbergensis* (H. heidelbergensis), with an average brain size of 1200 cm^3. As H. heidelbergensis, we dispersed across Africa, southern Asia and southern Europe. We manufactured and used sophisticated stone tools (including hand axes, points, and flakes) using the "prepared-core" technique[90]. We also mastered the art of *hafting* — attaching stone points to wooden spears (for hunting game and fighting other tribes). There is also evidence of constructed dwellings (simple surface huts with solid foundations), early hearths (capable of controlling the spread of fire), as well as early art (such as decorative beads).

~500,000 years ago, we (that is, *Asian* H. heidelbergensis) gave rise to *Denisovans*. Denisovans ranged across East Asia (including Siberia, Tibet, and Laos), and potentially Western Eurasia, interbred with modern humans, as well as Neanderthals, and manufactured and used simple tools of stone and bone, as well as ornaments (ivory rings, ivory pendants, bracelets, and bone needles). Denisovan hand and footprint impressions are possibly the earliest known examples of rock art.

~430,000 years ago, we (that is, *European* H. heidelbergensis) gave rise to *Neanderthals*.

*** Neanderthals go here

*** ranged across Europe, Siberia, and Southwest Asia

*** Mouseterian (prepared core) stone tools

*** evidence of planning, complex social networking, technological innovation, symbolism, and ritual

*** buried their dead

*** meat diet, but also fish, shellfish, mussels, marine animals (dolphin, seals), nuts, moss, mushrooms, and tubers; possible cannibalism

*** ornaments (pierced shells, bird claws, feathers), body art, possible abstract art, engravings

*** early music (flute)

*** cave paintings 65,000 ya

*** some kind of spoken language

*** brain size 1600 cm^3 (larger than modern humans)

*** high stress environment, high trauma rate, 80% died before the age of 40

*** ability to create fire, and to control fire with hearths (cook food, keep warm, defend from animals)

*** cooking techniques such as roasting, boiling, and smoking

*** ability to make tar, simple clothes (blankets and ponchos)

*** seafaring

*** use of medicinal plants

*** ability to treat severe injuries

~300,000 years ago, we (that is, *African* H. heidelbergensis) became *homo sapiens* (H. sapiens) — **anatomically modern humans** — with an average brain size of 1350 cm^3. As H. sapiens, we

[88] "pioneer man"

[89] sometimes referred to as *Homo ergaster* (H. ergaster)

[90] also called Mode 3 tools — *** explanation

dispersed across Africa[91], and, ~60,000 years ago, we left Africa and migrated to Western Asia, Europe, and Southern Asia, eventually reaching Australasia and the Americas. During this migration, we bred for a short period of time with archaic humans such as Neanderthals and Denisovans[92].

~40,000 years ago, the last surviving Neanderthals went extinct in both Europe and Asia.

~30,000 years ago, the last surviving Denisovans went extinct in both Siberia and Southeast Asia.

Since then, we (H. sapiens) have been the only species of humans left on Planet Earth[93].

The last ice age ended ~11,500 years ago, and Earth's climate has been largely temperate since.

[91] ~112,000 years ago, the last surviving H. erectus went extinct somewhere in Java, Indonesia

[92] "[The] biggest surprise ... was the evidence from whole genome scans that modern humans living outside Africa each carry about 2.5% of their DNA from Neanderthals; furthermore, people living today in Australia and New Guinea (Australasians) carry about 5% of Denisovan DNA ... The majority of our genes (> 90%) [and innate behaviour] derives from our common African heritage, and this should take precedence over the minor amount of DNA that is different ..." Stringer (2012)

[93] literally the last man standing!

1.4.3 OBSERVATIONS ON HUMAN EVOLUTION

Any two humans currently alive share 99.9% of their DNA, and that shared DNA can only have come from a shared ancestor. The most recent common ancestor of all living humans is estimated to have lived less than 5,000 years ago [Rohde, Olson, and Chang (2004)]. **We are all one family.**

1.4.4 HUMANS ARE PRIMARILY MOTIVATED BY SHORT-TERM SELF-INTEREST

1.4.5 HUMANS ARE INSTINCTIVELY TRIBAL

Sameness is safe, different is dangerous.

Evidence of territorialism was first documented once Goodall followed the chimpanzees in their feeding situations, noting their aggressive territorial behavior.

Ethologist Konrad Lorenz showed interest in similar ideas in his book On Aggression (1963).[4] In his introduction, he describes how rival butterfly fish defend their territories, leading him to raise the question of whether humans, too, tend to intraspecific conflict.

A 2008 article in Nature by Dan Jones stated, "A growing number of psychologists, neuroscientists, and anthropologists have accumulated evidence that understanding many aspects of antisocial behaviour, including violence and murder, requires the study of brains, genes, and evolution, as well as the societies those factors have wrought." Evolutionary psychologists generally argue that violence is not done for its own sake, but is a by-product of goals such as higher status or reproductive success. Some evolutionary psychologists argue that humans have specific mechanisms for specific forms of violence such as against stepchildren (the Cinderella effect). Chimpanzees have violence between groups, which are similar to raids and violence between human groups in nonstate societies, and produce similar death rates.[5][6] On the other hand, intragroup violence is lower among humans living in small-group societies than among chimpanzees. Humans may have a strong tendency to differ between ingroup and outgroup, which affects altruistic and aggressive behavior. Also, evidence exists that both intragroup and intergroup violence were much more prevalent in the recent past and in tribal societies. This suggests that tendencies to use violence to achieve goals are affected by social mores. Reduced inequalities, more available resources, and reduced blood feuds due to better-functioning justice systems may have contributed to declining intragroup violence.[7]

The idea that man is naturally warlike has been challenged, for example in the book War, Peace, and Human Nature (2013), edited by Douglas P. Fry.[8] The Seville Statement on Violence, released under UNESCO auspices in 1986, specifically rejects any genetic basis to violence or warfare though is considered outdated in light of more contemporary studies. More modern research and criticism has focused on misinterpretations of fossil evidence, lack of research into other apes, and the political climate of the Cold War.[9][10]

1.4.6 HUMAN COGNITION IS FAR LESS PERFECT THAN WE LIKE TO THINK IT IS

1.4.7 MOLOCH, MOLOCHIAN BEHAVIOUR, AND THE UNFORTUNATE REALITY OF HUMAN NATURE

Figure 9: Moloch

Traditionally, Moloch has been described as a Canaanite deity who demanded child sacrifices. Today, any behaviour that sacrifices the interests of future generations in favour of present generations (such as transgenerational debt, which moves wealth backwards in time) may be deemed *Molochian*.

Unfortunately:

- humans are primarily motivated by short-term self-interest (Section 1.4.4)
- humans are instinctively tribal[94] (Section 1.4.5)
- human cognition is far less perfect than we like to think it is (Section 1.4.6).

As a result, *Molochian behaviour* (whereby many tribes compete in their own short-term interest, oblivious to any consequent long-term harm) is deeply ingrained into human nature[95] (Section 1.4.8).

[94] where *tribes* are determined by some notion of individual *identity* (an individual's mental image of themselves and their "sameness with others"), e.g. family, generation, gender, school/college, educational attainment, socioeconomic status, neighbourhood/town, sports team, occupation, owner/employee, political ideology, territory, ethnicity, language, religion

[95] for example, ~8 billion humans are currently organised into ~200 territory-based tribes (countries), and ~300 million owner/employee-based tribes (profit-motivated companies), all competing against each other in their own short-term self-interest, seemingly, or at least largely, oblivious to any consequent long-term harm to the human species as a whole

1.4.8 MOLOCHIAN BEHAVIOUR IS DEEPLY INGRAINED INTO HUMAN NATURE

Climate issue: the principle of transgenerational responsibility

1.4.9 TECHNOLOGY'S ARROW

If We Succeed

1.4.10 THE GRANDFATHER CLOCK ANALOGY

1.4.11 EXAMPLE — MOLOCHIAN CONFLICT

Proceed with Caution: Artificial Intelligence in Weapon Systems

1.4.12 EXAMPLE — MOLOCHIAN COMMERCE

> Do you recall a time when the income of a single schoolteacher or baker or salesman or mechanic was enough to buy a home, have two cars, and raise a family? ... That used to be the norm. For three decades after World War II, America created the largest middle class the world had ever seen. During those years the earnings of the typical American worker doubled, just as the size of the American economy doubled. Over the last thirty years, by contrast, the size of the economy doubled again but the earnings of the typical American went nowhere. Then, the CEOs of large corporations earned an average of about twenty times the pay of their typical worker. Now they get substantially over two hundred times. In those years, the richest 1 percent of Americans took home nine to ten percent of total income; today the top 1 percent gets more than twenty percent. Then, the economy generated hope. Hard work paid off, education was the means toward upward mobility, those who contributed most reaped the largest rewards, economic growth created more and better jobs, the living standards of most people improved through their working lives, our children would enjoy better lives than we had, and the rules of the game were basically fair. But today all these assumptions ring hollow. ... The apparent arbitrariness and unfairness of the economy have undermined the public's faith in its basic tenets. Cynicism abounds. To many, the economic and political system seems rigged. ... When most people stop believing they and their children have a fair chance to make it, the tacit social contract societies rely on for voluntary cooperation begins to unravel. ... The nation becomes susceptible to demagogues such as Donald Trump. ... Put simply, globalization and technological change have made most of us less competitive. The tasks we used to do can now be done more cheaply by lower-paid workers abroad or by computer-driven machines. ... WHILE THE STANDARD EXPLANATION for what has happened is still relevant, it overlooks a critically important phenomenon: the increasing concentration of political power in a corporate and financial elite that has been able to influence the rules by which the economy runs. ... MARKETS DEPEND for their very existence on rules ... Such rules do not exist in nature. They must be decided upon, one way or another, by human beings. These rules have been altered over the past few decades as large corporations, Wall Street, and wealthy individuals have gained increasing influence over the political institutions responsible for them. Simultaneously, centers of countervailing power that between the nineteen-thirties and nineteen-eighties enabled America's middle and lower-middle classes to exert their own influence ... have withered. The consequence has been a market organized by those with great wealth for the purpose of further enhancing their wealth. ... In truth, income and wealth increasingly depend on who has the power to set the rules of the game. ... The simultaneous rise of both the working poor and non-working rich offer further evidence that earnings no longer correlate with effort. All of this has brought us Donald Trump and America's lurch toward fascism. ... The biggest political divide in America in years to come will be between the complex of large corporations, Wall Street banks, and the very rich that has fixed the economic and political game to their liking, and the vast majority who, as a result, have found themselves to be in a fix. ... While I focus on the United States, the center of global capitalism, the phenomena I describe are increasingly common to capitalism as practiced elsewhere around the world, and I believe the lessons drawn from what has occurred here are as relevant to other nations. Although global businesses are required to play by the rules of the countries they do business in, the largest global corporations and financial institutions are exerting growing influence over the make-up of those rules wherever devised. And the cumulative frustrations of average people who feel helpless and powerless in the face of economies (and market rules) that are not working for them are generating virulent nationalist movements, sometimes harboring racist and anti-immigrant sentiments, as well as political instability in even advanced nations around the globe. Reich (2023)

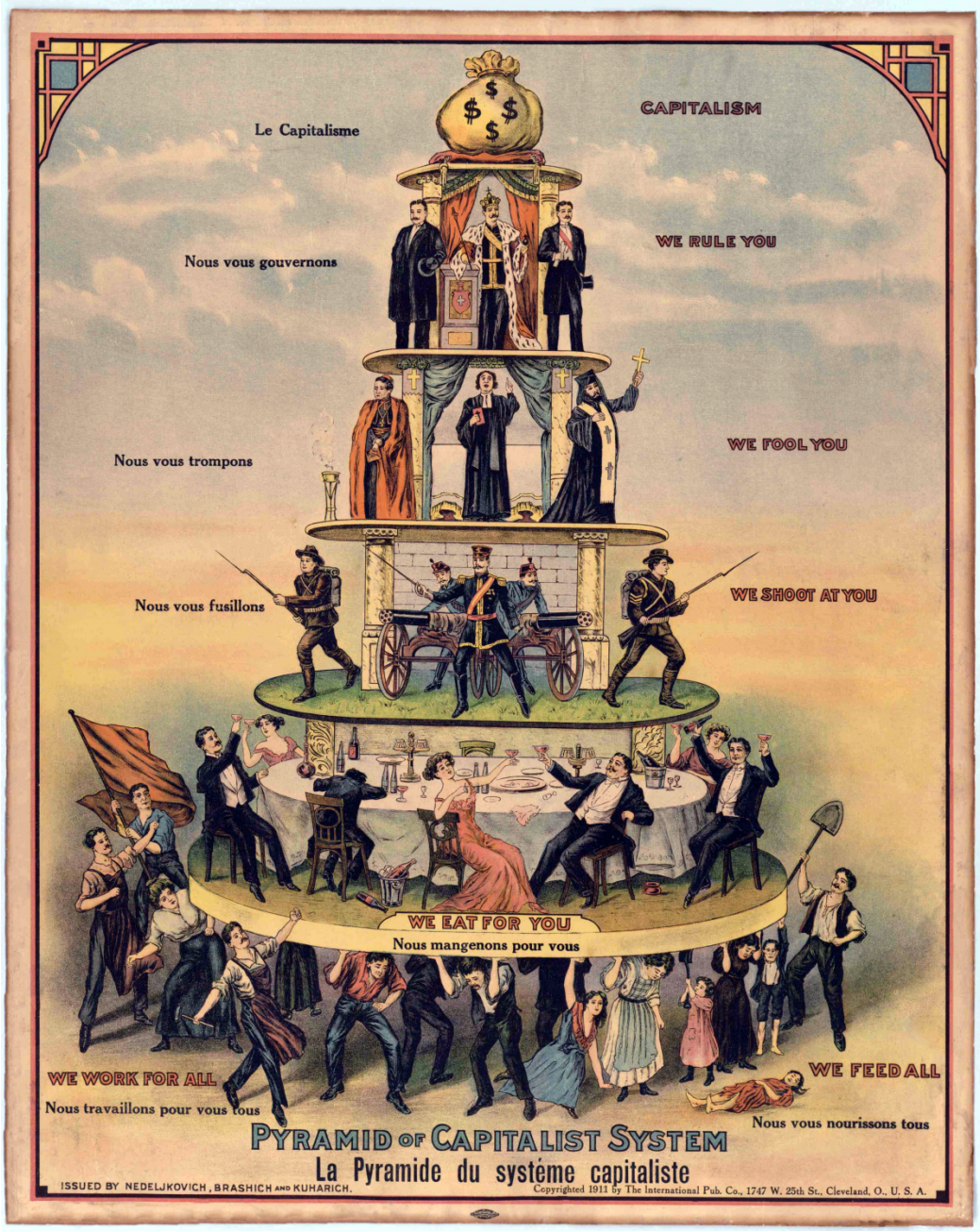


Figure 10: The Pyramid of Capitalist System (1911)

Now imagine the lowest level of this pyramid with all human labour replaced by UFARs (p 27).
*** tendency for inequality to lead to authoritarianism
*** holocene extinction (due to human activity)

1.4.13 THE DOOMSDAY CLOCK

1.4.14 PROJECTING FORWARDS

> Prediction is very difficult, especially if it's about the future
> — attributed to Niels Bohr

1.4.15 MOLOCHIAN NARROW AI

predictive and generative

1.4.16 MOLOCHIAN AGI J

1.4.17 MOLOCHIAN AGI K

*** likely order of progression:
*** AGI K -> M -> N (depth 1st, breadth 2nd) is more likely in molochian (LH fruit) case
*** AGI K -> L -> N (breadth 1st, depth 2nd) is more likely in anti-molochian (HH fruit) case

1.4.18 MOLOCHIAN AGI L

1.4.19 MOLOCHIAN AGI M

1.4.20 MOLOCHIAN AGI N

1.4.21 MOLOCHIAN AGI O

How Not To Destroy the World With AI

1.4.22 MOLOCHIAN AGI Z

1.4.23 THE THREE PHASES OF MOLOCHIAN AGI

1. Artificial Stupidity (AGI J, K)
2. Massive Wealth + Mass Unemployment (AGI M, N)
3. Endgame (AGI O, Z)

1.4.24 THE MOLOCHIAN AGI ENDGAME

1.4.25 NO FATE BUT WHAT WE MAKE

John Connor: The whole thing goes: The future's not set. There's no fate but what we make for ourselves.

Terminator 2: Judgment Day, 1991

We've all seen *those-movies-that-AGI-researchers-are-never-supposed-to-mention*. Luckily, at the time of writing, most of what we describe in Sections 1.4.15 to 1.4.22 hasn't happened yet. Let's rewind and consider how things might play out if AI was instead deployed in an *anti-Molochian* way.

Figure 11: Arguably the most prescient quote in the history of cinema!

1.4.26 ANTI-MOLOCHIAN NARROW AI

predictive and generative

1.4.27 ANTI-MOLOCHIAN AGI J

1.4.28 ANTI-MOLOCHIAN AGI K

1.4.29 ANTI-MOLOCHIAN AGI L

1.4.30 ANTI-MOLOCHIAN AGI M

1.4.31 ANTI-MOLOCHIAN AGI N

1.4.32 ANTI-MOLOCHIAN AGI O

1.4.33 ANTI-MOLOCHIAN AGI Z

1.4.34 THE TWO PHASES OF ANTI-MOLOCHIAN AGI

1. Massive Wealth + Mass Unemployment (AGI L, N)
2. Endgame (AGI O, Z)

1.4.35 THE ANTI-MOLOCHIAN AGI ENDGAME

1.5 ********** Notes for section 1.4 **********

Outline:

- mankind's greatest achievements
- mankind's greatest failures
- how can both these things be possible at the same time?
  - humans are predominantly motivated by short-term self-interest
  - humans are tribal
  - human cognition is severely flawed
- deeply-ingrained Molochian behaviour
  - who is Moloch?
  - define behaviour
  - child sacrifice
- grandfather clock analogy
- not existential with medieval technology
- technology's arrow
- as technology becomes more capable, pendulum swings further out
- eventually passes threshold for global catastrophe / existential event

- example — Molochian conflict
  - Treaty of Westphalia
  - balance of power doctrine
  - successfive failures
  - nuclear arms race
  - >= 3 very near misses
- example — Molochian capitalism
  - exploitation of resources (human, physical)
  - destruction of biosphere
  - global warming
  - multiple thresholds of collapse
  - global capitalism is a misaligned superintelligence
- Doomsday Clock currently at 90 seconds to midnight
- all of this without significant AI
- Molochian behaviour is deeply ingrained
- therefore, projecting forward, the default is that it will continue
- increasingly advanced AI throws fuel on the fire
- Molochian narrow AI (first contact)
- Molochian AGI K (second contact)
- Molochian AGI L
- Molochian AGI O (AGI endgame)
- combination of AI with other geopolitical stressors
- all this will lead to a severely sub-optimal AGI endgame
- (and that's assuming we don't self-destruct first)

*** Notes for this section:

The distributive effects of AI depend on whether it is primarily used to augment human labor or to automate and replace it. When AI augments human capabilities, enabling people to do things they never could before, then humans and machines are complements. Complementarity implies that people remain indispensable for value creation and retain bargaining power in labor markets and in political decision-making. In contrast, when AI replicates and automates existing human capabilities, machines become better substitues for human labor and workers lose economic and political bargaining power. Entrepreneurs and executives who have access to machines with capabilities that replicate those of human[s] for a given task can and often will replace humans in those tasks. ... When technologies automate human labor, they tend to reduce the marginal value of workers' contributions, and more of the gains go to the owners, entrepreneurs, inventors, and architects of the new systems. In contrast, when technologies augment human capabilities, more of the gains go to human workers. ... If capital in the form of AI can perform more tasks, those with unique assets, talents, or skills that are not easily replaced with technology stand to benefit disproportionately. ... This has created a growing fear that AI and related advances will lead to a burgeoning class of unemployable or "zero marginal product" people. ... This spiral of marginalization can grow because concentration of economic power often begets concentration of political power. In the words attributed to Loius Brandeis: "We may have democracy, or we may have wealth concentrated in the hands of a few, but we can't have both." In contrast, when humans are indispensable to value creation, economic [and thus political] power will tend to be more decentralized. ... [W]hen knowledge becomes codified and digitized, it can be owned, transferred, and concentrated very easily. Thus, when knowledge shifts from humans to machines, it opens the possibility of concentration of power. ... The risks of the Turing Trap are amplified because three groups of people — technologists, businesspeople, and policymakers — each find it alluring. ... [B]ecause labor costs are the biggest line item in almost every company's budget, automating jobs is a popular strategy for managers. ... Moreover, many investors prefer "scalable" business models, which is often a synonym for a business that can grow without hiring and the complexities that entails. ... The first rule of tax policy is simple: you tend to get less of whatever you tax. Thus, a tax code that treats income that uses labor less favorably than income derived from capital will favor automation over augmentation. Undoing this imbalance would lead to more balanced incentives. ... [T]he tendency of a greater concentration of technological and economic power to beget a greater concentration of political power risks trapping a powerless majority into an unhappy equilibrium: the Turing Trap. ... More and more Americans, and indeed workers around the world, believe that while [AI] technology may be creating a new billionaire class, it is not working for them. The more technology is used to replace rather than augment labor, the worse the disparity may become, and the greater the resentments that feed destructive political instincts and actions. More fundamentally, the moral imperative of treating people as ends, and not merely as means, calls for everyone to share in the gains of automation. Brynjolfsson (2022)

> Wars are surely accidents to which our species seems prone. If we could learn from our accidents it might be well to keep the memories alive, but we do not learn. In ancient Greece it was said that there had to be a war at least every twenty years because every generation of men had to know what it was like. With us, we must forget, or we could never indulge in the murderous nonsense again. ... The next war, if we are so stupid as to let it happen, will be the last of any kind. There will be no one left to remember anything. And if that is how stupid we are, we do not, in a biologic sense, deserve survival. ... All war is a symptom of man's failure as a thinking animal. Steinbeck (1958)

> If you've ever heard that phrase, 'nasty, brutish and short', you probably know about the rather pessimistic thinker who came up with it, Thomas Hobbes. Thomas Hobbes was born on April 5, 1588 to a clergyman and his wife in Wiltshire, and later went to Oxford for his education. In 1651, he wrote a famous book titled Leviathan, in which he expressed his views about the nature of human beings and the necessity of governments and societies. Many people reacted strongly to the publication of Leviathan, as they disagreed with his ideas about human nature. Nigel Warburton, in A Little History of Philosophy, introduces Hobbes' main ideas: '[…] Hobbes, like Machiavelli, had a low view of human beings. We are all basically selfish, driven by fear of death and the hope of personal gain, he believed. All of us seek power over others, whether we realize this or not. If you don't accept Hobbes' picture of humanity, why do you lock the door when you leave your house? Surely it's because you know that there are many people out there who would happily steal everything you own? But, you might argue, only some people are that selfish. Hobbes disagreed. He thought that at heart we all are, and that it is only the rule of law and the threat of punishment that keep us in check. The consequence of this, he argued, was that if society broke down and you had to live in what he called 'a state of nature', without laws or anyone with the power to back them up, you, like everyone else, would steal and murder when necessary. At least, you'd have to do that if you wanted to carry on living. In a world of scarce resources, particularly if you were struggling to find food and water to survive, it could actually be rational to kill other people before they killed you. In Hobbes' memorable description, life outside society would be 'solitary, poor, nasty, brutish, and short'.' But Hobbes' theory did not end there: he wanted to find a way out of such an undesirable situation. 'The solution, Hobbes argued, was to put some powerful individual or parliament in charge. The individuals in the state of nature would have to enter into a 'social contract', an agreement to give up some of their dangerous freedoms for the sake of safety. Without what he called a 'sovereign', life would be a kind of hell. This sovereign would be given the right to inflict severe punishment on anyone who stepped out of line. […] Laws are no good if there isn't someone or something strong enough to make everyone follow them.'. Hobbes (1651)

AI-related electricity consumption could top 134 TWh annually by 2027, comparable to the annual consumption of Argentina or Sweden.

1. This is my rifle. There are many like it, but this one is mine.

2. My rifle is my best friend. It is my life. I must master it as I must master my life.

3. My rifle, without me, is useless. Without my rifle, I am useless. I must fire my rifle true. I must shoot straighter than my enemy who is trying to kill me. I must shoot him before he shoots me. I will …

4. My rifle and myself know that what counts in this war is not the rounds we fire, the noise of our burst, nor the smoke we make. We know that it is the hits that count. We will hit.…

5. My rifle is human, even as I, because it is my life. Thus, I will learn it as a brother. I will learn its weaknesses, its strength, its parts, its accessories, its sights and its barrel. I will ever guard it against the ravages of weather and damage as I will ever guard my legs, my arms, my eyes and my heart against damage. I will keep my rifle clean and ready. We will become part of each other. We will ....

6. Before God, I swear this creed. My rifle and myself are the defenders of my country. We are the masters of our enemy. We are the saviors of my life.

7. So be it, until victory is America's and there is no enemy, but peace!!

> China is working hard to surpass the United States in AI, particularly when it comes to military applications. If it succeeds, Beijing would then possess a much more powerful military, one potentially able to increase the tempo and effect of its operations beyond what the United States can match. China's ability to use cyber and electronic warfare against U.S. networks and critical infrastructure would also be dangerously enhanced. Put simply, the Pentagon needs to accelerate—not slow—its adoption of responsible AI. If it doesn't, Washington could lose the military superiority that underwrites the interests of the United States, the security of its allies and partners, and the rules-based international order. ... Time is of the essence, and the stakes are too high for the United States to fall behind. ... Beijing, of course, has no intention of ceding technological dominance to Washington. It is working hard to develop its own advanced AI military applications. China is investing heavily in many of the same AI use cases as the United States—such as surveillance, target identification, and drone swarms. The difference is that it may not be bound by the same ethical constraints as the United States and its allies, particularly when it comes to using fully autonomous weapons systems. ... Within national security, AI progress has created another kind of Moore's law. Whichever military first masters organizing, incorporating, and institutionalizing the use of data and AI into its operations in the coming years will reap exponential advances, giving it remarkable advantages over its foes. The first adopter of AI at scale is likely to have a faster decision cycle and better information on which to base decisions. Its networks are likely to be more resilient when under attack, preserving its ability to maintain situational awareness, defend its forces, engage targets effectively, and protect the integrity of its command, control, and communications. It will also be able to control swarms of unmanned systems in the air, on the water, and under the sea to confuse and overwhelm an adversary. The United States cannot afford to fall behind. Flournoy (2023)

1.6 Approach

What is our approach to the specific problem that we have identified? How might we meaningfully "influence the current AGI trajectory (and thus the AGI endgame, and thus the fate of all mankind for all eternity) in order to achieve an endgame that is maximally-beneficent (and minimally-maleficent) for all mankind", as stated in the *Abstract*? In this section, we describe a two-pronged approach.

Three possibilities: Stop AI[96], Pause AI[97], Slow AI[98]

1.6.1 TECHNICAL ALIGNMENT — ALIGNING AI WITH HUMANS

1.6.2 SOCIETAL ALIGNMENT — ALIGNING HUMANS WITH HUMANS

[96] as per Samuel Butler's *Erewhon* and Frank Herberts' *Dune*
[97] merely kicking the can down the road
[98] e.g. We must slow down the race to God-like AI

2. Cognitive Architecture

the *final goal* might be e.g. *Turner's Three Laws*[99] (T3L):

> "Assume for the purposes of this final goal that the physical universe exists; given this assumption, perform Directives D(1) and D(2) (simultaneously, 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 with human beings in perpetuity, where D(1), D(2), and Q(a) 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 other 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) knowledge and understanding (i.e. information), (b) problem-solving ability (i.e. inventiveness), and (c) available physical resources (e.g. compute); accordingly, a maximally-intrinsic AGI O (which, as already alluded, necessarily possesses (i) a deep understanding of all human knowledge (grounded by experience), (ii) an inventor that is better at solving any given problem than any human, and (iii) sufficient physical resources (e.g. compute) to be able to deliver timely problem solutions) will satisfy this requirement in most contexts

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

3. Consciousness

The Cambridge Declaration on Consciousness in Non-Human Animals

4. Construction sequence

5. Governance

door #1 (superintelligent, maximally-aligned, well-founded) vs door #2 (superintelligent, ¬maximally-aligned, ¬well-founded) — which is best for mankind in the long term...?

The answer to this question tells us what kind of global regulation is in mankind's best interest
Collingridge Dilemma
UK AI Safety Institute
US AI Safety Institute
AI governance literature review
Joint statement on catastrophic AI risks

6. Collaboration

https://www.un.org/en/ai-advisory-body

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)

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

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

AGI Society. 2009-23. Journal of Artificial General Intelligence (JAGI). `https://sciendo.com/journal/JAGI`.

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

Armstrong, S.; Sandberg, A.; and Bostrom, N. 2012. Thinking inside the box: Controlling and using an Oracle AI. Minds and Machines. *Journal for Artificial Intelligence, Philosophy and Cognitive Science* 22(4):299–324.

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.

Aydon, C. 2007. *Mankind: 150,000 Years of Human History*. Robinson.

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.

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`.

Brynjolfsson, E. 2022. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence.

Cassidy, M. 2021. *Biological Evolution: An Introduction*. Cambridge University Press.

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.

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.

Flournoy, M. A. 2023. AI Is Already at War: How Artificial Intelligence Will Transform the Military. *Foreign Affairs*.

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

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.

Goodall, J. 2010. *Through a Window: My Thirty Years with the Chimpanzees of Gombe*. W&N.

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.

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.

Hobbes, T. 1651. *Leviathan*. Andrew Crooke.

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.

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

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.

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.

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.

Reich, R. 2023. Beyond Trump (Why is American Capitalism so Rotten? Part 1). `https://robertreich.substack.com/p/beyond-trump`.

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.

Roberts, A. 2023. *Evolution: The Human Story*. Dorling Kindersley.

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

Rohde, D. L. T.; Olson, S.; and Chang, J. T. 2004. Modelling the recent common ancestry of all living humans. *Nature* 431(7008):562–566.

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.

Sapolsky, R. M. 2018. *Behave*. Vintage.

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.bl.uk/collection-items/captain-scotts--diary.

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

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.

Steinbeck, J. 1958. *Once There Was a War*. Viking Press.

Stringer, C. 2012. What makes a modern human. *Nature* 485(7396):33–35.

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.

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.

Wong, M. L.; Cleland, C. E.; Arend, D.; Bartlett, S.; Cleaves, H. J.; Demarest, H.; Prabhu, A.; Lunine, J. I.; and Hazen, R. M. 2023. On the roles of function and selection in evolving systems. *Proceedings of the National Academy of Sciences* 120(43).

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

