Artificial General Intelligence – induction, deduction, abduction

Aaron Turner, January 2021

We examine the AGI architecture at the heart of the BigMother project.

Truth

It is, sadly, literally impossible for any intelligent entity (either human or machine) to observe the universe and determine **with absolute certainty** what is true and what isn't. Literally impossible.

In the final analysis, an intelligent entity can never be 100% certain that the physical universe it perceives actually exists (as it might just be some kind of illusion). Similarly, and contrary to Descartes, an entity can't even be certain of its own existence (for example, there might be some valid mechanism that logically implies that you don't exist, in any possible way, despite your own perception to the contrary, that you (indeed, all humans) lack the cognitive capacity to conceive; that possibility, however tenuous, is sufficient to be less than 100% certain of your own existence).

Assuming that both a sole universe and a given entity exists, the closest that entity can ever get to knowing “the truth” would be for it to have made every possible observation of the universe at every point in time from the birth of the universe (believed to have occurred some 13.8 billion years ago) to the present, simultaneously at every point in space. Even this would not necessarily constitute complete knowledge, but such a vast corpus of information would nevertheless give the entity in question a pretty good idea of what the true state of the universe was, is, and to some extent will be. In reality, of course, most humans (for example) will have made only an infinitesimal fraction of all possible observations over the course of considerably less than a hundred years, limited in physical space to that tiny fraction of the planet Earth that they visit in their lifetimes.

As well as the information practicably available to an entity being woefully incomplete, it will often be very noisy. In general, any physical observation may be corrupted by either random noise (fluctuations around the true value), systemic bias (a consistent offset from the true value), or both. Examples would be not being able to distinguish a dog from a fox due to poor lighting, not being able to properly hear what someone is saying due to background chatter, or consistently measuring 2 kg heavier than actuality due to your bathroom scales being improperly zeroed.

Belief

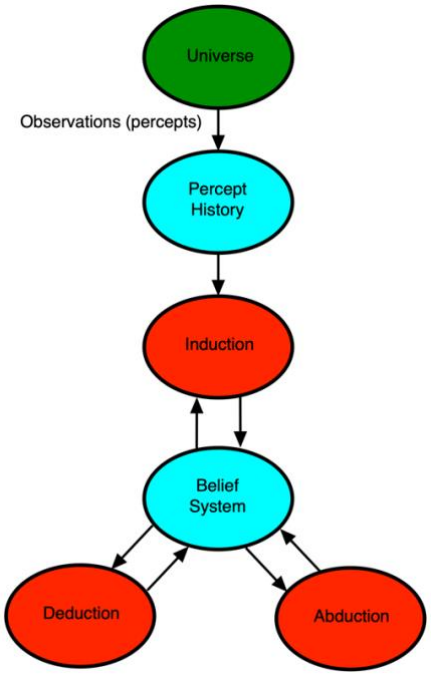
In the presence of these metaphysical considerations, and the absence of complete information, the best an entity can ever hope to achieve in actual practice is to synthesise an **approximation** to the truth (effectively a **guess**) derived from a necessarily finite set of noisy observations. This approximation constitutes the entity’s **belief system**, or **internal model** of the external universe.

Considered purely logically (and deliberately ignoring how, say, the human brain might attempt to do so), the process of synthesising reliable beliefs from a necessarily finite and incomplete set of noisy observations is fraught with difficulty, as exemplified by Plato’s allegory of the cave. Plato imagines a group of people who have spent their entire lives in a cave, facing, and chained to, a blank wall, such that the only thing they can ever directly observe are the shadows cast onto the wall from a fire behind them. These shadows are their only clues as to the nature of the universe.

By analogy, everything that we, as humans, have ever directly experienced in our entire lives via our senses are just **shadows on the wall**, and it is from this infinitesimal sliver of incomplete and often noisy information that we must each attempt to construct our own personal belief systems.

Cognition

Conceptually (i.e. logically), the process by which beliefs are formed may be imagined as follows:



Simply stated, each of the three logical cognitive processes (induction, deduction, and abduction) takes one or more **existing beliefs** as input, and generates one or more **new beliefs** as output. For our purposes, we assume that the belief system is initially devoid of beliefs pertaining to the physical universe, but has been initialised with a number of axioms, definitions, and theorems of mathematics. As a first approximation, we can think of a belief as being a **declarative statement** (one that is either true or false, despite the practical difficulties of determining exactly what these things mean). More concretely, we might think of a belief as being a **wff** (well-formed formula), or (in BigMother's case) a **theorem**, of some foundational formal system such as First-Order Logic with Equality (FOLEQ), first extended with the axioms of set theory, and then further extended with a toolkit of mathematical definitions (built on top of that set theory), including a definition of probability, in order that beliefs may then be associated with some notion of "degree of belief".

Induction

The **inductive** cognitive process (**induction**) is the only route by which beliefs pertaining to the physical universe may be added to the belief system, which it does by continually observing the universe via the entity's built-in senses / input devices [delete as appropriate]. Each observation generates a **percept** (timestamped packet of information) which is added to the entity's **percept history** (the collection of all observations of the universe that the entity has made in its lifetime).

Once the percept history contains sufficient observations, induction can start to do its job, which essentially boils down to detecting **patterns** (**regularity**) in the data. Conceptually, every time a pattern is detected in the data, a **belief** (theorem) corresponding to a declarative statement of the

form “pattern X exists at location Y (in the percept history) with qualification Z” may be added to the belief system. “Qualification Z” is required because the pattern match will often be less than perfect. For example, because the raw percept data is potentially noisy, sometimes a pattern might only be determined to be present in the data with a certain probability (e.g. “I want to say it’s a dog, but it could also be a fox”). In extreme cases, the pattern that is present in actuality might be so distorted by noise in the percept data that it’s mischaracterised (e.g. hearing “four candles” instead of “fork handles”). Similarly, it might be the case that a “phantom” pattern is detected in the data that is not present in actuality (known as a false positive, or Type I error, e.g. perceiving a dog when no dog is present), or alternatively it might be the case that no pattern (at all) is detected in the data when one is present in actuality (known as a false negative, or Type II error, e.g. not perceiving a dog, when a dog is present). Finally, a pattern might only be partially present in the percept data (a partial match), such as when only the front half a dog is visible.

Note that the patterns detected by induction are not necessarily limited to percepts originating from a single sense / input device. Patterns may be detected across percepts from the same source (such as within visual data), from two or more different sources (multi-modal patterns, such as a correlation between visual and audio data), across percepts sampled at the same time (such as stereo vision), or across percepts sampled at different times (temporal patterns, such as motion in visual data). In order to facilitate multi-modal pattern matching, it is advantageous for percepts to use the same generic representation (such as a labelled digraph), irrespective of their source. In other words, as far as the inductive cognitive process is concerned, it’s just data.

Once sufficient observations have generated sufficient data (percepts) that multiple patterns have been detected, and corresponding beliefs added to the belief system, induction really starts to get interesting, because at this point the beliefs that have been synthesised so far may themselves be viewed as data (ideally encoded using exactly the same internal representation as percepts), and the induction process can start to recognise any patterns in those beliefs, each of which then generates a new, higher level (i.e. **more abstract**) belief, which may then itself be viewed as data at the next level up. And on and on the process continues, ad infinitum. As long as the entity continues to make new observations, more and more patterns will be detected at each level, and higher and higher level (i.e. more abstract) beliefs will be synthesised. In principle, “I wandered lonely as a cloud” (and, in general, any sentiment or idea) may be deconstructed into its lower level patterns, continuing recursively until bottoming out at the level of raw percepts.

As a significant refinement to this process, once an entity has sufficient data, it can start to make empirical probability estimates (e.g. “so far I’ve seen 1729 dogs, and 1727 of them have had four legs, whereas 2 have had only three legs; therefore, given my accumulated experience to date, there’s an (estimated) 99.884% probability that any particular dog will have four legs”). More generally, an entity can estimate (marginal, joint, conditional) probability distributions in respect of the population of all possible observations given its finite sample of empirical observations.

Given sufficient experience (i.e. a sufficiently large body of observations of the universe pushed through induction as above), there is, in principle, no knowledge that is beyond an AGI’s reach. Because beliefs are declarative statements, or, more concretely, wffs/theorems of formal logic, the beliefs synthesised by **induction** are immediately amenable to **deduction** and **abduction**.

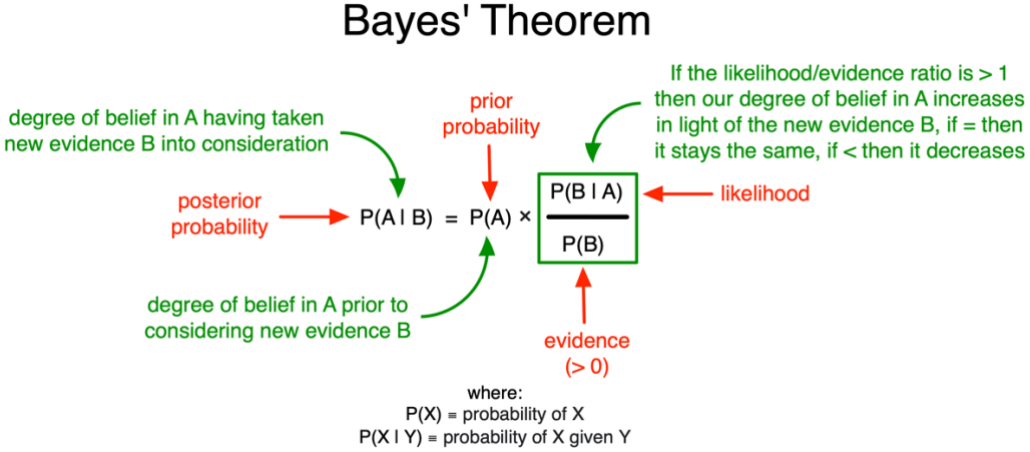
Deduction

If the universe was perfectly random then the only thing an intelligent entity would ever see when observing it would be random noise, with no structure to it, i.e. **no regularity**, and so the induction process would never detect any patterns. Luckily, our universe does have structure, i.e. it's not perfectly random, which means that it can be **predicted** – not necessarily with perfect accuracy, because randomness is still present, but nevertheless with better accuracy than random chance.

This is vitally important, because most, arguably all, biologically-evolved intelligent entities (such as humans) have at least some kind of implicit motivation, objective, or top-level **goal** (such as avoiding pain, avoiding being eaten, securing adequate food, water, and shelter, engaging in happy fun time, etc), and an ability to make **better-than-random predictions** about the external universe from its internal model of the universe is a vital prerequisite for **any goal-directed entity**.

The **deductive** logical cognitive process (**deduction**) relates pre-synthesised beliefs B_1, …, B_n with their **logical consequences** C_1, … C_m. So, for example, if an intelligent entity has partially matched a dog in its percept history then belief B might have the form "pattern [dog D] exists at location Y (in the percept history) with qualification [50% match, only two legs visible]", from which (given the accumulated prior experience alluded above) the entity will be able to **infer** new belief C: "there's an (estimated) 99.884% probability that [dog D] has four legs" (in AI, this is known as **forward chaining**). Similarly, the **abductive** logical cognitive process (described below) will often be facilitated in its operation if the deductive logical cognitive process is able to determine whether or not some **conjecture** C (such as "there's a > 99% probability that [dog D] has four legs") is a logical consequence of B_1, …, B_n (this is known as **backward chaining**, or **theorem proving**). In either case, the net effect of deduction is to **predict** (i.e. reveal) **latent information** (C_1, … C_m) that isn't **explicitly** present in the set of beliefs B_1, …, B_n, but is instead **logically implied** by it.

Because the data in an entity's percept history is incomplete and noisy, any beliefs derived from observation of the external universe are necessarily **probabilistic** in nature, and thus so are any deductive inferences involving these externally-derived beliefs. If the entity in question has made sufficient observations to be able to reliably estimate empirical probability distributions, then it may possess sufficient statistical information to be able to use **Bayes' Theorem**, thereby allowing certain probabilistic beliefs about the external universe to be updated in the light of new evidence:



Abduction

The **abductive** cognitive process (**abduction**) is often described as "reverse deduction". In its simplest form, if an intelligent entity has made some observation leading to some belief C, then abduction asks the question "what **hypothetical** belief H would **explain** the observation (that led to belief C)?" In logical terms, we wish to find some belief (declarative statement, or wff) H such that H **implies** C. For example, having observed a three-legged dog D, H might be "D was born that way", "D was hit by a car", "D bit his own leg off", or a myriad other possibilities, each of which **potentially** explains why D only has three legs. Because so many (often an infinite number of) possible hypotheses potentially explain the observation in question, any hypothesis H is at best a **guess**. Thus, **absent further investigation**, hypothesis H is merely an **unverified possibility**.

Any hypothesis H therefore needs to be **assessed**, rather than simply **assumed**. A hypothesis H (which seeks to explain some observation leading to belief C) is (a) **valid** if and only if (IFF) the wff "H implies C" is a theorem, (b) **likely** IFF its associated probability exceeds a suitably high threshold, and (c) **plausible** IFF it is logically consistent with those pre-existing beliefs that are held with unimpeachably high confidence. Ideally, any hypothesis should be further assessed using a combination of (d) **statistical analysis**, and/or (e) established **scientific method**, such as conducting experiments designed to verify any predictions derived from a hypothesis. Given sufficient knowledge (of statistics, scientific method, and the universe) derived from experience (observation) via induction (as described above), there is, in principle, no level of evidence-based critical thought (and therefore critical assessment of hypotheses) that is beyond an AGI's reach.

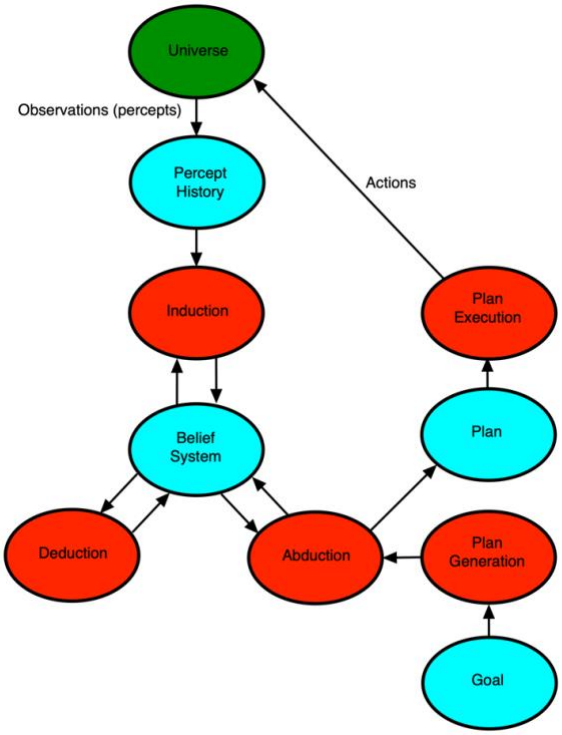
Abduction may be generalised to problems of the form "find (term) t such that t has property P". Standard abduction as described above is logically equivalent to "find t such that t is a declarative statement and t implies C", and theorem-proving is equivalent to "find t such that t is a proof that C is a logical consequence of B_1, …, B_n". The generalised form of abduction may be used to define literally **any creative problem**, for example "find t such that t is a C++ program to calculate Fast Fourier Transforms", "find t such that t is a musical composition and, when performed, t will have a positive emotional impact on humans that enjoy the musical genre known as jazz", "find t such that t is a cure for cancer", or "find t such that t is an effective algorithm that, when executed, will strive to (a) determine actual human preferences, and (b) maximise the extent to which actual human preferences are realised". An implementation of generalised abduction might attempt to find a suitable t via an algorithm known as state space search, perhaps accelerated by a neural net (and/or various other methods, such as massively parallel processing, custom hardware, etc); a necessary prerequisite would be that the entity had already acquired sufficient domain-specific / common sense (i.e. world) knowledge, for example by comprehensively observing the universe via induction. As already alluded above, state space search will often need to perform deduction.

Closing the sense-effect loop

This combination of induction, deduction, and abduction is believed (at least by the author) to be **cognitively complete**, meaning that any cognitive process, in any intelligent entity (either human or machine), however physically implemented (either biologically or electronically) is **functionally equivalent** to some arbitrarily-complex sequence of inductions, deductions, and abductions.

The inductive-deductive-abductive architecture so far described, however, merely (a) observes the external universe, thereby gradually accumulating a percept history of necessarily incomplete and noisy data, and then (b) attempts to construct (from that percept history, via induction) an approximate internal model of the universe in order to (c) facilitate better-than-random prediction about the universe (via deduction), and (d) solve problems (pertaining to the universe) expressed in terms of generalised abduction. In AGI terms, the mechanism so far described is an **oracle**, of which we can potentially ask questions, but not yet an autonomous goal-directed **agent** capable of pursuing assigned objectives without requiring continuous micro-management or supervision.

In order to upgrade our oracle to an agent, we add **plan generation and execution**, as follows:



This might look like a lot of extra work, but actually it isn't, as we already have most of what we need. Overall, either implicitly or explicitly, the agent has some kind of top-level **goal** (objective), from which the plan generation process (taking into consideration everything the entity believes about the universe, as represented by its belief system) synthesises (and continuously updates) a **plan** (essentially analogous to a computer program), which the plan execution process then executes (analogously to how a computer program would be), which then results (via the entity's output devices) in external **actions** that impinge upon, and thereby affect, the external universe. Clearly, however, the plan generation process can simply hand off the goal to the abduction cognitive process, which generates a plan on its behalf, and thereby does all the heavy lifting.

Thus we have closed the **sense-effect loop**: (conceptually, logically, functionally) the intelligent entity, however concretely implemented, now interacts with the universe in pursuit of its goal.

(Note: in an AGI context, **extreme care** must be taken to ensure that the machine's goal, and overall behaviour, are **forever maximally-aligned** with mankind's goals, values, and objectives!)