Chapter 2
Background

It's all these black boxes you can't open—see how each spends most of its time trying to defeat the other.

Knuth [368]

Recent years have seen an explosion in academic, industrial, and popular interest in AI, as exemplified by machine learning and primarily driven by the widely-reported successes of deep- and reinforcement learning (e.g. [314, 315, 351]). Deep learning is essentially predicated on the notion that, with a sufficiently large training set, the statistical correlations captured by training will actually be causal [310]. However, in the absence of convergence theorems to support this, it remains a hypothesis. Indeed, insofar as there *is* evidence, it increasingly indicates to the contrary, since the application of enormous volumes of computational effort has still failed to deliver models with the generalization capability of an infant. There is accordingly increasing discussion about what further conceptual or practical insights might be required [57]. At the time of writing, the very definition of deep learning is in flux, with one Turing Award laureate defining it as "a way to try to make machines intelligent by allowing computers to learn from examples"[1] and another as "differentiable programming".[2] We argue in the following that deep learning is highly unlikely to yield intelligence, at the very least while it equates intelligence with "solving a regression problem". Specifically, we claim that it is necessary to adopt a fundamentally different perspective on the construction of inferences from observations, and that this is in accordance with a fundamental revolution in the philosophy of science: Karl Popper's celebrated solution to 'The Problem of Induction' [268].

[1] https://blogs.microsoft.com/ai/a-conversation-ai-pioneer-yoshua-bengio.

[2] https://www.facebook.com/yann.lecun/posts/10155003011462143.

© The Author(s) 2022

J. Swan et al., *The Road to General Intelligence*,
Studies in Computational Intelligence 1049,
https://doi.org/10.1007/978-3-031-08020-3_2

In subsequent chapters, we first describe the significant challenges for machine learning. We then argue that current approaches are unlikely to be able to address them unless the scope of the learning framework is considerably widened, and that it will ultimately be necessary for this framework to be both situated and to support reflective reasoning. We compare our proposed conceptual framework for learning with that of reinforcement learning, with respect to the features necessarily associated with general intelligence. We propose a roadmap towards general intelligence, progressing via the notion of *work on command* to the property of *semantic closure*. This property is described fully in Chap. 7, but in précis, it equips an agent with the ability to determine causality and induce hierarchical representations in a manner that is absent from traditional machine learning approaches.

2.1 What we Mean by General Intelligence

As convincingly argued by Wang [357]: before any in-depth discussion of roadmaps and obstacles, we must define the desired destination of general intelligence. Legg and Hutter [194] give a comprehensive and insightful tour of various definitions of intelligence. They distill many of these into the following observations:

> *Intelligence is not the ability to deal with a fully known environment, but rather the ability to deal with some range of possibilities which cannot be wholly anticipated. What is important then is that the individual is able to quickly learn and adapt so as to perform as well as possible over a wide range of environments, situations, tasks and problems.*

These observations along with others culminate in the mathematical formalism of *universal intelligence*. This defines the intelligence of an agent as being equal to the average of the returns (sum of rewards) it can obtain across all possible environments, weighted by the complexity of those environments.

The preceding definition is problematic in that it assumes the existence of an *a priori* reward function. In natural organisms, any such reward function is assumed to be provided by a combination of innate and cultural mechanisms. For artificial systems, the implied complexity of a constructed reward function becomes a concern. In the extreme, if the reward function were to be completely arbitrary, then it effectively characterizes the environment as pure noise, with no useful features for a learner to exploit. Conversely, once the reward function is structured so as to reflect regularities in the environment, then at least some (and potentially a great deal) of the purported intelligence of the agent is actually provided by the reward function itself.

In subsequent sections, we argue that some of the issues with ML are actually an artifact of this kind of 'narrow framing of the problem', in which (1) human expertise is required to represent each specific problem in a manner amenable to ML and (2) this framing then only makes a highly impoverished form of feedback available the learner, typically in the form of a scalar numeric reward. We argue that it is necessary

to replace black box reward functions with richer representations that are suitable for reflective reasoning.

In pursuit of greater generality than can be provided by an *a priori* reward function, we must therefore adopt the wholly pragmatic perspective of the following value proposition.

The Value Proposition for General Intelligence

For all practical purposes, general intelligence is a necessary property of a system which:

- Performs *work on command*
  i.e., responds with tolerable latency to dynamic changes in goal specification and environmental conditions.
- Scales to real-world concerns.
- Respects safety constraints.
- Is explainable and auditable.

2.2 Science as Extended Mind

There is clearly a huge gap between the software which enables facial or gait recognition and the yet-to-be-realized technology which will allow safe and trustworthy autonomous vehicles or factories. One can likewise consider the reality gap between audio-activated digital assistants and fully-fledged household robots. There exist countless other examples of roles that current AI techniques are incapable of fulfilling. In roles where humans are currently irreplaceable, what *traits* enable them to meet the demands of these roles? The generality of human intelligence is evident in many ways:

- Humans can handle multiple objectives simultaneously and can typically order activities so as to meet these objectives relatively efficiently.
- Humans can learn skills without forgetting those previously learned. They can also make efficient use of related skills to bootstrap their learning process and minimize this effort.
- Humans can explain their decision-making in terms of relevant causal factors and 'locally consistent' frameworks of thinking, which means that a recipient of the explanation (perhaps also their subsequent self) can understand, verify, and possibly rectify the steps taken to reach conclusions.
- Humans can be told what is desired directly as a goal rather than needing to iteratively try behavior in the hope of optimizing some sampled metric.
- Humans can be told what is forbidden and/or constrained and they can avoid such situations without needing to physically interact with (i.e. 'sample') the environment, assuming relevant grounded world knowledge.

- Humans can gracefully adjust their cognitive resource usage between perception, action, and learning rather than having rigid boundaries between them.[3]
- Humans can operate in multi-agent settings, mostly through being able to effectively model other agents' trajectories based on their perceived intentions and behavioral patterns.
- Humans can do all of the above in the real world, perhaps with a curriculum, but not needing a high-fidelity resettable/reversible simulation within which to learn.

For the purposes of this work, the above list of traits will be considered as a set of necessary emergent capabilities of general intelligence. With this motivation, we claim that a system which exhibits these traits must satisfy the requirements summarized in Fig. 1.1.

While the human mind is the most immediate exemplar for general intelligence, we believe there are strong reasons to consider that the *scientific method* is better suited to provide a template for its implementation. As Rodney Brooks has famously observed [37], insights obtained via mental introspection might cause us to be deeply misled about the nature of intelligence. In contrast, the adoption of the scientific method yields falsifiable statements about the physical world. This can be seen as providing an 'extended mind' [50]—an externalized artifact with *verifiable properties* that can directly inform the design of general intelligence architectures. Given the inevitable concerns about 'AI alignment',[4] such verifiability is of particular importance in obtaining measures of *safety*. Hence, we believe that the path to general intelligence (at the very least, in a form capable of respecting safety concerns) lies in the attempt to automate the scientific method, from the perspective of an embodied reasoner with real-world concerns of deadlines and resource availability.

Recent years have seen increasing emphasis on causality in machine learning. Causality is essential for building reasoning systems as it is a stronger criterion than merely statistical correlation. Originally having been convincingly argued for by Pearl [254], the relevance of causality to AI has been since agreed upon by Schölkopf, Bengio, and others. One of the key ideas is the 'ladder of causality,' which is framed as inference situated on three 'rungs': the observational, interventional, and counterfactual settings. Statistical learning from fixed datasets operates solely the observational rung: training on data generated only by an external process which the model does not affect. Interventions imply the ability to set values of certain variables despite the natural external processes in order to generate informative data; for example, double-blind experiment design with control groups. The most demanding yet powerful application of causality is counterfactual reasoning, where inferences are drawn based on variable values which were never observed but generated through interventions in a model; for example, alternate history timelines.

Pearl also introduced the *Structural Causal Model* (SCM), which is a directed acyclic graph structure specifically designed to enable users to operate on all three rungs of the ladder of causality. A key idea in the SCM formalism is that dependencies

[3] Whilst not necessarily under conscious control, nonetheless a property of human cognition overall.

[4] The quest for confidence that a general intelligence won't attempt to turn everything into paperclips [30].

between variables are framed as probabilistic *functions* rather than simply statistical dependence. This is better aligned with a physical interpretation of observations, namely that they are caused by physical processes over time. The distributions of a set of variables X_i are given by the formula:

$$X_i = f_i(\mathrm{PA}_i, U_i), \quad (i = 1, \ldots, n)$$

where PA_i are the parent nodes of X_i. The functions are probabilistic due to U_i, which are exogenous noise variables which are jointly independent of one another. If there were dependencies, they could be explained by forming yet more causal relationships (as per the common cause principle), and so noise must be modeled as independent.

Interventions in SCMs are defined as (temporarily) setting f_i to be a constant. Importantly, the distributions of the parent nodes are unaffected by interventions on children since their relationships are effectively severed, which is different from standard Bayesian networks. Instances of the latter only denote conditional independence relationships as undirected graphs, and so dependence between nodes persists even if the value of one is set. The ability to perform interventions also allows for principled counterfactual reasoning in SCMs. If we had observed some value for a node X_i, we can use abduction to estimate the value of U_i, and then after intervening on its parents PA_i, re-apply the observed exogenous noise in order to produce a counterfactual inference. There are ubiquitous problems which require counterfactual reasoning that are consequently intractable for purely statistical models [254]. Given their apparent completeness for causal modeling, the modern problem of causal discovery involves deducing the topology of an SCM which can accurately describe the system.

Despite widespread interest in the use of SCMs, it is vital to appreciate that, within scientific practice, causality is best understood as being only part of a contextualized process of situated, bidirectional inference. Hence we take the deeper view of science as the construction of statements which provide a concise and consistent description of *possible worlds*. As recently observed by David Deutsch:

> *Finding causal theories is necessary but not sufficient. We need* explanatory *theories. "Mosquitos cause malaria" is essential and useful but only 1% of the way to understanding and curing malaria.*

The essence of the scientific method is, of course, the interleaving of problem formulation, hypotheses generation, experimentation, and analysis of results. Hence, a core aspect of the proposed approach is the requirement for a reflective expression language. Reflection is the property that statements can themselves be treated as data, meaning that hypotheses about knowledge in the language can be evaluated as first-class objects. This becomes salient for the process of *hypotheses generation*.

Concretely, for our purposes hypotheses are some (sub)graph of inferences in the system's transition model. Mappings from sensor inputs to effectors (or in the opposite direction, in the case of abductive inference) are just specific fragments of this overall model, starting or terminating in appropriately designated sensor or effector dimensions. Naturally, it is desired that the only hypotheses that are entertained by

the system are those which (1) actually describe a possible world and (2) are relevant to the task at hand. Considered from a 'traditional symbolist' perspective, the latter is of course equivalent to the well-known 'Frame Problem' [219], which, as discussed in subsequent chapters, is increasingly understood to have been an artefact of coarse-grained and disembodied inference.

As we discuss in detail in Chaps. 7 and 9, it also follows that any reasonable candidate architecture for encoding knowledge for general intelligence will be *compositional*, so as to obtain a semantics for compound hypotheses. Another essential property is the notion of *strong typing*, which enables inheritance/taxonomy and the explicit denotation of goals and constraints as regions of a (prospectively open-ended) state space. These properties jointly enable structured updates to working knowledge that retain the self-reinforcing nature of a scientific theory [89].

Finally, the modern scientific method requires that all working knowledge should be *in principle falsifiable* via empirical observation. Hence, we include as a requirement to our approach that the base symbols of the expression language must include denotations which are grounded in this way. Causal modeling also stipulates the ability to intervene directly in the environment to learn the effects of one's own agency. Thus, the system and representation language must both support primitives for interacting with the environment.

We claim that it will not be possible (or economical) to automate human labor in the general case until AI also possesses these properties. As such, this is the context in which we will highlight the challenges and shortcomings of deep learning, reinforcement learning, and other existing AI approaches. To place contemporary approaches in the appropriate context, we proceed via a brief historical recapitulation of the rise and fall of the traditional symbolist approach.

2.3 The Death of 'Good Old-Fashioned AI'

The key figures at the inaugural AI conference at Dartmouth [220] were split across the nascent symbolist and connectionist divide, with Simon, Newell, and McCarthy in the former camp, and Shannon, Minsky, and Rochester in the latter [174]. However, the symbolist approach became the prevailing one, not least because of the widespread confusion surrounding the solvability of the XOR problem by perceptrons [226]. The following decades saw concerted effort in symbolic AI. Many of the languages used to construct AI originated from the synthesis of procedural and logical programming styles, respectively exemplified by LISP and resolution theorem provers. Hewitt's 'PLANNER' language [142] was a hybrid of sorts, being able to procedurally interpret logical sentences using both forward and backward chaining. It was used to construct SHRDLU [366] which was hailed as a major demonstration of natural language understanding. This inspired other projects such as CYC [197], an ongoing attempt to create a comprehensive ontology and knowledge base that seeks to capture 'common-sense knowledge'. Less ambitious and more successful were the various expert system projects that started in the 1960s and became

prevalent in the following two decades. These included the MYCIN expert system for diagnosing infectious disease [349], Dendral for identifying unknown organic molecules [38] and other well-known systems such Prospector [135]. Collectively, such systems became known as 'Good Old-Fashioned AI' (GOFAI) [137].

Common Attributes and Roadblocks

In general, GOFAI fits the template of a knowledge-based system. Such systems may be decomposed into two components: a knowledge base and an inference engine which answers queries and/or enhances the knowledge base by applying inference rules to the current state of the knowledge base. These systems typically required users to create the symbolic primitives and inference rules a priori. It gradually became understood that such information was difficult to obtain—the so-called 'knowledge elicitation bottleneck'.

Two other key longstanding GOFAI problems are the Qualification Problem and the Frame Problem [219], both of which contributed to the perception of scalability issues. The Qualification Problem is concerned with the preconditions needed for a logical deduction to be valid. The Frame Problem is concerned with the difficulty of fully specifying conditions related to invariants of state space transformation.[5] Another key obstacle to scalability is the 'cognitive cycle' of the 'sense–think–act' paradigm [240], in which it is assumed that the world evolves in lockstep with the system. This quickly became a formidable obstacle for early projects such as the General Problem Solver [238] and PLANNER [142]. At that time, computation was also far more expensive than it now is—together, these two things meant that GOFAI was destined for a reckoning.

The End of GOFAI

((The prospect of general intelligence using such rule-based systems was not highly appraised by observers. In the early 1970s, the Lighthill Report [202] lead to a drastic reduction in AI research funding by the UK government and DARPA cut funding to academic AI research in the USA. In the late 1980s, the United States' Strategic Computing Initiative, which was essentially formed to participate in an AI/computing race with Japan, cut funding for new AI research as its leaders realized that the effort would not produce the full machine intelligence it desired. Simultaneously, the market for large-scale expert system hardware collapsed and more affordable general-purpose workstations took over. These developments formed part of what is now colloquially known as the 'AI Winter'.

In hindsight, many of these challenges and the accompanying demise of GOFAI could be said to be a function of the hardware of the time. Computing power and mem-

[5] It was eventually concluded that solutions exist to these problems, including default logics [347] and answer set programming [201].

ory were obviously more expensive and software design decisions (such as deciding between the use of LISP or C/C++) had a correspondingly disproportionate impact on what could be computed in practice. A number of companies built upon LISP-based expert systems (such as Symbolics, LISP Machines Inc., Thinking Machines Corporation, and Lucid Inc.) went bankrupt. Ambitious undertakings were common, such as the Fifth-Generation Computer Systems (FGCS) project in Japan during the 1980s. The massive investment in building highly parallel computing systems was ultimately in vain as simpler, more general architectures such as workstations from Sun Microsystems and Intel x86 machines became favored for such roles. Some of those forward-looking ideas have been reinvented in the early 21st century, such as the emphasis on highly parallel programming from FGCS now commonplace in general-purpose GPU programming with CUDA and OpenCL.

The Problem of the 'Sense–Think–Act' Loop

Regardless of technological advances, the GOFAI paradigm still does not present a viable path to general intelligence. For the architectures relying on ungrounded knowledge representation, there is no prospect of deploying them to address tasks in the real world of complex and noisy data streams. More fundamentally, the absence of grounding precludes the understanding of causal relationships of the real world—a core aspect of operationalizing the scientific method. Even if a GOFAI system were hypothetically to achieve symbol grounding, there would still be a fatal flaw: GOFAI never matured sufficiently to escape the scalability problem inherent in the 'sense–think–act' loop. As the system's body of knowledge grows, the time required to make plans and predictions must also increase. As engineers put it, the system 'lags behind the plant' and faces two options: either to deliver correct action plans but too late, or to deliver on time plans that are incorrect [241]. This issue arises essentially from the *synchronous* coupling of the agent and its environment, i.e., the latter is expected to wait politely until the agent completes its deliberations. Technically speaking, synchronicity means that the agent computes in zero time from the environment's perspective.

Machine learning has failed to acknowledge the significance of this problem and has even adopted GOFAI's synchronous coupling as one of the fundamentals of reinforcement learning; see Chap. 3. For now, computation is scaling at a rate that can sustain the synchronous abstractions used in large-scale projects (see Sect. 5.2). For lower-level routines (such as reactively handling sensory streams of data at a fixed frequency) this may suffice. On the other hand, there are certainly aspects to cognition which are slower, more deliberative and explicitly logical, and this is where the 'sense–think–act' approach breaks. Some believe that given more resources and innovations in model architectures, deep learning may be able to encode knowledge effectively enough to empower this sort of cognition and meet the requirements for general intelligence. In the next chapters, we shall see why this, in fact, cannot be the case.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.