Proceedings of Machine Learning Research 159:5–27, 2021 International Workshop on Self-Supervised Learning

The 'Explanation Hypothesis' in General Self-Supervised Learning

Kristinn R. Thórisson THORISSON@RU.IS
Center for Analysis and Design of Intelligent Agents, Reykjavik Univ., Iceland
& Icelandic Institute for Intelligent Machines, Reykjavik, Iceland

Editors: K. R. Thórisson and Paul Robertson

Abstract

Self-supervised learning is the ability of an agent to improve its own performance, with respect to one or more goals related to one or more phenomena, without outside help from a teacher or other external aid tailored to the agent's learning progress. A *general* learner's learning process is not limited to a strict set of topics, tasks, or domains. Self-supervised and general learning machines are still in the early stages of development, as are learning machines that can explain their own knowledge, goals, actions, and reasoning. Research on explanation proper has to date been largely limited to the field of philosophy of science. In this paper I present the hypothesis that general self-supervised learning requires (a particular kind of) explanation generation, and review some key arguments for and against it. Named the *explanation hypothesis* (ExH), the claim rests on three main pillars. First, that any good explanation of a phenomenon requires reference to relations between sub-parts of that phenomenon, as well as to its context (other phenomena and their parts), especially (but not only) causal relations. Second, that self-supervised general learning of a new phenomenon requires (a kind of) *bootstrapping*, and that this – and subsequent improvement on the initial knowledge thus produced – relies on *reasoning processes*. Third, that general self-supervised learning relies on *reification* of prior *knowledge* and *knowledge-generation processes*, which can only be implemented through appropriate *reflection* mechanisms, whereby current knowledge and prior learning progress is available for explicit inspection by the learning system itself, to be analyzed for use in future learning. The claim thus construed has several important implications for the implementation of general machine intelligence, including that it will neither be achieved without reflection (meta-cognition) nor explicit representation of causal relations, and that internal explanation generation must be a fundamental principle of their operation.

Keywords: Self-Supervised Learning, Explanation, General Machine Intelligence, Cumulative Learning, Machine Learning, Seed Programming, Knowledge Representation, Autonomy, Generality, Autonomous Generality

1. Introduction

'Learning' is the ability of an agent to improve its own performance with respect to one or more (high- or low-level) goals related to one or more (simple or complex) phenomena; self-supervised learning is the ability to learn without outside help from a teacher. The primary benefit of self-supervised learning for artificial intelligence (AI) systems is the reduction in requirements for manual specification of learning targets and a lowered need for overseeing the learning process, which is another way of saying that, at its ultimate level of success, general self-supervised learning is about the possibility of realizing full autonomy in knowledge acquisition.[1] Learning comes in many forms and is dependent on many things other than a controller's learning machinery, including what kind regularities a world presents, what data is available at any point in time, what sensors a learner can access to measure its own progress and changes in the world (and how well it knows those sensors), and what knowledge it comes endowed with upon deployment (or birth).

The kind of learning we are looking at here is general self-supervised learning, also called *cumulative learning* (Thórisson et al. (2019); Thórisson and Talbot (2018)), in worlds where the number of possible elements and valid (time-dependent) arrangements is vastly greater than any learner could ever enumerate. Machines capable of general self-supervised learning in such environments are still mostly at the theoretical or blueprint stage of development (cf. Thórisson et al. (2014)). There is reason to believe that over the next few decades such systems will see the light of day and become common in industrial automation, as they promise increased resilience and flexibility compared to contemporary approaches.

The concept of 'explanation' – its role, nature, and generation – has recently come into focus in AI, in large part due to the recent increase of contemporary AI for automating a variety of industrial purposes (cf. Miller (2019)), including risk assessment (e.g. insurance claim handling, job hiring), visually-guided control (e.g. autopilots in automobiles), speech recognition (e.g. dictation, translation), and many other. A primary technology for this purpose has been artificial neural networks (ANNs). ANN technology can be applied in situations where no other prior technology would work; after proper training, ANNs can transform large amounts of data into actionable knowledge, via a large network of classification functions (Wang and Li (2016)), to enable automatic control for tasks involving complex real-world data, where coding rules by hand would be prohibitively expensive, take prohibitively long, or simply impossible. Rather than being driven by a desire for explanation capabilities per se, the request for explanation in large ANN systems is due to their inherent opaqueness: explaining how a deployed ANN came to its conclusions, concisely and coherently, at adequate and variable levels of detail and abstraction, has proven difficult. This may, in no small part, be due to their inability to represent causal relations. As of yet, these systems are not well-positioned for routine explanation in practical contexts.

While valid after-the-fact explanation is certainly a worthy research goal were it to allow investigation of technologies whose behavior begs – for whatever reason – to be explained. But given that that the field of AI concerns itself with automation, automating the explanation process itself should perhaps be higher on researcher's to-do list.[2] The operation of a machine with hand-coded knowledge has, other things being equal, much less reason for explaining itself than a learning machine whose knowledge is – due to having been learned autonomously – largely unknown to its designers. As a result of being assigned numerous tasks of critical importance in society, the inability of modern automation to provide logical arguments for why they do what they do – has become a hot topic: people whose work, life, or actions are affected by them want to know why they took certain actions and not others, especially when these actions are perceived as inappropriate; their designers want to

[1] We consider "full autonomy" to be measured from the moment a controller is deployed into its target operating environment.

[2] For this, ANNs may even be even less well-positioned than they are for third-party explanation as it calls for continuous, on-line, incremental learning, which in turn requires goal-directed control.

understand the logic behind their behavior. Accessing this at desired levels of abstraction – in the same way that explanations of events allow humans to asses the sources of mistakes, chains of events, etc. – is of utmost importance when designing automation for complex systems. Their designers might also want to implement preventative measures for such systems *before* they are deployed, for the same reasons. This calls for an ability to provide *valid explanations*.[3] This is the underlying premise for the topic of the present paper.

No existing theory of explanation lends itself as an obvious basis for building AI systems, and learning machines that can automatically explain their own knowledge, goals, actions and reasoning, are in short supply. Research on explanation proper has been largely conducted in the philosophy of science, with virtually no coherent treatment in the field of artificial intelligence. Contemporary AI automation methods are largely confined to well-defined (narrow) tasks, the topic of generality and autonomy in learning are seldom put front and center, which is certainly part of the reason why explanation in AI has focused primarily on building *explainable* systems, as opposed to systems that can *automatically explain themselves*. We take one step further by proposing that explanation is a foundational process in general autonomous learning; explaining thoughts and actions *to others* are thus not a topic here, but rather, internal explanations necessary for learning.

The topic of this paper is the *explanation hypothesis* (ExH), whose main premise is that general self-supervised learning necessarily depends on explanation. This view on the role of explanation goes counter to the view of explanation as a social phenomenon (cf. Miller (2019)), where its main purpose is to convey information from an explainer to an explainee. The argument behind the hypothesis – as presented here – does not rest on some technical, unusual, or particular definitions of the key relevant terms (e.g. 'general', 'learning', 'explanation', etc.) but quite the contrary, we try to stay close to these concepts as commonly used in the vernacular; the way we use these terms here should not be a surprise. That being said, to clarify this further (and in more detail than can be done with a generic reference to the vernacular), the next section goes into detail on the particular assumptions and definitions behind key terms on which the hypothesis relies. Thereafter we look at some relevant related work, which helps provide a further context for the meaning of key concepts used. Then we present the hypothesis in a compressed form, followed by an expanded view of the ideas on which it rests. Lastly, we discuss some arguments for and against it and some of its interesting implications.

2. Key Terms & Definitions

Worlds. In the context of the present paper, a world W refers to the full scope of where tasks are performed, consisting of a set of variables ($\mathcal{V}$), behaving in accordance with a set of spatio-temporal transformation functions ($\mathcal{F}$) operating on these over time, resulting in (time-dependent) relations ($\Re$) between them; $W = \langle \mathcal{V}, \mathcal{F}, \Re, \mathcal{S}_0, \mathcal{T} \rangle$, where $\mathcal{S}_0$ is a world's initial state and $\mathcal{T}$ is time. The world has a clock-on-the-wall, meaning that time progresses independently of any intelligent agent's activities in it. The result are patterns on numerous

[3] We consider the validity of explanations to be concretely measurable through analysis and experimentation—if not in practice, then at least in principle. Our use of the concept of 'valid explanations' might thus be more accurately described as "practically (and/or theoretically) validatable explanations."

scales of organization, containing regularities at multiple levels of detail, in accordance with the values of the variables as determined by the transformations over time. The prime example of this kind of dynamic world is the physical one. Such highly complex worlds present infinite variety that nevertheless follows strict rules; we call them infinite worlds.

Agents, Embodiment, & Experience. For an agent in such a world, only a small subset of the variables are observable ($\mathcal{V}_o$) at any point in time, and only some are manipulable ($\mathcal{V}_m$). An agent is a source of agency that can change the values of manipulable variables and measure the values of observable ones, through the operations and interactions of its controller, which schedules actions and measurements, and its body, which consists of transducers.[4] We assume relatively fixed transducers[5] with a limited scope, that is, they are inadequate for measuring the full spatio-temporal range of the world on any dimension (i.e. they downsample). A set of measurements presented to an agent's controller, via its transducers, and the internal operations that it is able to reflect on, i.e. that are accessible to the learning system itself via reflection, is referred to as the agent's experience. The learner's experience is time-sensitive because the learning must keep track of time; as experience accumulates the learner's knowledge state thus progresses by default. Given these assumptions, a learning agent is guaranteed to never encounter all possible combinations of any set of relations between any set of (types of) elements (and can never even be sure it has done so, even if its transducers cannot tell the difference and make them *seem* identical) as we assume that the size of the world (i.e. the number of elements and relations) approaches infinity, $\|\Re\| \to \infty$, and the set of observable variables is a vastly smaller set than the complete set of variables in the world, $\|\mathcal{V}_o\| \lll \|\mathcal{V}\|$.

Environments. An environment is a *particular* subset of the world (e.g. 'my kitchen,' 'highlands of Iceland,' etc.). In these, variables have specific bindings. An environment may belong to a family of environments (e.g. "kitchens in the Western world"), in which case variables do not have specific bindings (for instance, my particular kitchen contains no sauce pan, but the *family* of kitchens may). To ground the above concepts of world and environment in something practical related to learning, it may be useful to think of the smallest difference that can be measured by our learning agent's transducers.

Domains. A domain, as used here, is the subset of worlds containing particular types of environments, as defined above. A domain harbors a potential for a variety of types of tasks with a variety of forms (also called task family). In the vernacular, the term 'domain' is used for a set of environmental constraints with semi-homogeneous transformation functions, elements, and types of relations. Examples include 'indoor environments,' 'a forest,' 'urban environments,' etc. The fact that its boundaries are not crisp is a feature, not a bug: unless otherwise noted, when we use the term in this paper we typically mean to include its full scope of potential meaning, from a narrow to expansive interpretation.

Phenomenon. Any useful grouping of a subset of spatio-temporal patterns experienced by an agent in an environment may be called a phenomenon. A phenomenon Φ in the world is any grouping of variables ($\mathcal{V}_\Phi$) and relations ($\Re_\Phi$) that we choose to group as such; $\Phi = \langle \mathcal{V}_\Phi, \Re_\Phi | \mathcal{V}_\Phi \subseteq \mathcal{V}_\mathcal{W} \wedge \Re_\Phi \subseteq \Re_\mathcal{W} \rangle$. It consists of elements $\{\varphi_1 \ldots \varphi_{\|\Phi\|} \in \Phi\}$ that may

[4] By 'transducers' is meant devices that transform energy from one state to another, whether for output (e.g. motors or muscles) or input (i.e. sensing or measuring).

[5] That is, we assume that these have a given stable 'simplest mode' of operation, which allows learning to bootstrap from low-level principles, e.g. correlational data.

themselves consist of other phenomena, variables, and relations $\Re_\Phi$ (causal, mereological, etc.) and which most often are closely related spatio-temporally. $\Re_\Phi$ couples elements of Φ with each other, and with those of other phenomena in the world (Bieger and Thórisson (2017), Thórisson et al. (2016)), and can be partitioned into inward facing relations $\Re_\Phi^{in}$ between element pairs $\varphi_i, \varphi_j \in \Phi$ and outward facing relations $\Re_\Phi^{out}$ between element pairs $\varphi_i \in \Phi$ and $\beta_j \in W \backslash \Phi$.

Tasks. A task is a set of one or more goals and constraints, pertaining to an environment (or a domain which can be guaranteed to have actual variable bindings for the goal description), that is sufficiently detailed to be assigned to an agent to be performed. An assignable task has always a bound valid start interval and end time, an implicit maximum energy (total and per time unit), and typically also a set of constraints (negative goals, i.e. things to be avoided). Comparing the goals to measurements of the variables it references in the environment is sufficient for determining whether a task has been completed.

Goals. By goal is meant a compact specification of a set of states and procedures that reference sets of variables in a domain and is specific enough to serve as verification that a task has been performed successfully. Goals, under this view, are thus also sufficient to verify learning progress (by comparing a learner's performance on comparable goals in similar situations over time). Any time the performance of a task is to be verified, as well as learning progress towards that task, a goal must be explicitly articulated (whether overtly or only covertly).

Goals may form a hierarchy, where a single super-goal is composed of two or more sub-goals that describe particular constraints (e.g. sequence) for achieving it. For instance, travelling between A and B may include choosing several modes of transportation, based on the kind of terrain between A and B (boats for oceans, cars for roads, etc.). Planning such travel involves identifying useful sub-goals to sequence the events that lead to a successful progression from A to B (first by boat, then by airplane, ...). Any real-world task may have any number of associated negative goals (constraints) as part of the specification; this, along with available resources, determine which sub-goals will make sense for an agent agent performing a particular task.

We can identify the goals of machines, and their ability to achieve those goals, in light of the reason and purposes for which they were built. Such goals are persistent top-level goals or drives. A single machine will typically have many such goals (for instance, the purpose of an automobile is not only to get people from their chosen A to B (on land) but also to keep them safe, within a particular temperature range, able to control the car, etc.; the purpose of an internal combustion engine is to provide thrust; keeping a constant speed is the goal of cruise control; the goal(s) of the cruise control are the sub-goals of a car's passengers).

Goals come in two forms, explicit and implicit. To be part of task's definition (e.g. do the dishes), a goal must be explicit, that is, explicitly refer to variables, goals and sub-goals of importance (e.g. "all plates and cups must be clean before noon today"). Explicit goals are easily encoded (and communicated) in a compressed form (like the preceding example), assuming that the receiver of the goal can decompress such a description (bind variables and relate the task description to phenomena in the target environment). Implicit goals are those which are not explicit and not even necessarily obvious from the design of a machine or process, but can be inferred from observations of a machine's behavior. For a machine whose purpose is unknown, capturing its core *tendency/ies* in a compact high-level description, based on its observed behaviors, can serve as an ad-hoc explication of an otherwise implicit goal. This is called a reified goal; the process of making implicit information explicit is called reification.

Only explicit goals – i.e. goals that are given a particular compact description – can be directly compared and contrasted and treated explicitly in an information system; for instance, systematically modified, put on hold and returned to later, explicitly abandoned, etc.

Learning & Representation. The process by which particular experience is transformed into a representation of some form in an agent's memory is called 'learning.' The storage requires some medium (e.g. neurons) and representation format (information encoding). In the present paper we are only concerned with format. Generally speaking, a representation of something is, by definition, a model of that thing (Conant and Ashby (1970)). Strictly speaking, the models[6] created in this way model relationships between correlated sensations, which include sensations of actions under an agent's own volition, as well as anything else including (some of) its internal processes—we will discuss this latter point in the section on implications. Here we always talk of a controller's measurements (i.e. experience); a model of this experience may be used as a stand-in of the thing experienced, and for convenience, typically is.[7] Here we assume that models of separate experiences are related to each other in various ways (for instance, if the learner can keep track of time, the models will be related through some form of temporal ordering). Regularity in the world may be exploited in the modeling process to increase learning speed, by increasing and improving such generalizations over time. The full set of models that an agent may harbor at any time is called the agent's 'knowledge.'

General Learning. A 'general learning machine' is a machine whose learning mechanisms can be successfully applied to multiple task, domains, situations, environments, and worlds (but not necessarily with equal efficiency). Generality, in this view, is a gradient that can be specified and measured by quantifying *variety* related to a learner's target task(s), along one or more dimensions (independent of the scale used, whether absolute, relative, nominal, ordinal, discrete or continuous), and comparing the ability of two or more learners on these dimensions (cf. Thórisson (2020)).

Self-Supervised Learning. A controller that is capable of self-supervised learning can learn without the help of a teacher. By 'teacher' is meant any external aid specifically targeted to increase a particular agent's quality of learning, i.e. produce a net increase in its speed, retention, comprehension, and/or scope. The kind of self-supervised learning we are interested in here is *autonomous* general learning (AGL),[8] which combines self-supervision with general learning, as defined above.

[6] By 'model' we mean a composite information structure containing a phenomenon's invariants, elements with their relations and constraints, which together can be used to answer questions about a phenomenon, as a whole or its parts, including predictions, goal achievement, explanations, and re-creation (Bieger and Thórisson (2017); Thórisson et al. (2016)).

[7] In other words, although the only evidence anyone has of anything in the world is their experience of it (assuming Descartes was right that the only thing we can be sure of is that – to paraphrase – "I *am* because I *think*"), it is often more practical to take a third-person view here and talk directly about the phenomena hypothesized to be the cause for the experience.

[8] We consider 'general self-supervised learning' and 'autonomous general learning' to be synonymous.

Cumulative Learning. As used here, the concept of cumulative learning has been addressed in the AI literature to some extent, *but its many necessary-but-not-sufficient features* have invariably been addressed *in isolation.*[9] *Always-on* learning has for instance variously appeared under the headings 'lifelong', 'perpetual', 'never-ending', 'incremental', 'online', and 'continual' learning (Fontenla-Romero et al. (2013); Mitchell et al. (2018); Silver et al. (2013); Zhan and Taylor (2015); Zhang (2018)), most of which only have partial overlap with our use. We return to this topic a bit more at the end of the Related Work section below.

Logical Argument is the systematic application of rules to a set of premises to derive new information and meta-information, for the purpose of providing a judgement along (one or more) dimensions of comparison between (two or more) knowledge elements, where the data may be (one or more) measurements, rules, premises, etc. relevant to (spatio-temporal) subsets of a World, Domain, Task-Environment, or Phenomena. In short, the application of reasoning to a set of (derived or given) premises to produce explicit statements or models of target phenomena, and whose usefulness (validity) can be ascertained through empirical evaluation in the world they reference. It may employ mutual exclusion (based on non-axiomatic, defeasible principles (Wang and Awan (2011), Pollock (2010))) to generate a reasoned explanation of such analysis. The new knowledge produced is 'meta' because its subject is other knowledge or information. The application of logic is an 'argument' because its output "argues in favor" of one or more interpretations of a set of comparisons over one or more others. In the context of the present paper, the interpretations have to do with the usefulness of the information in question for a particular purpose, vis à vis the usefulness of learning particular things about particular phenomena in particular circumstances for particular ends (i.e. in light of set goals), and the ability to carry that information to different situations, contexts, and goals.

3. Related Work

Halpern and Pearl (2005b) describe a theory of explanation based around causal structural diagrams. While these causal diagrams are compositional, the work does not address the task of autonomous creation of such causal diagrams (which is needed to make an autonomous self-supervised learner). Nevertheless, their work is highly relevant to the implications of ExH. Halpern and Pearl (2005b:891) define explanation as follows:

> *The role of explanation is to provide the information needed to establish causation.* [...] *we view an explanation as a fact that is not known for certain but, if found to be true, would constitute a genuine cause of the explanandum, regardless of the agent's initial uncertainty. Thus, what counts as an explanation depends on what you already know and, naturally, the definition of explanation is relative to the agent's epistemic state* [...]

Compatible with Halpern and Pearl's theory, Thórisson et al. (2016) proposed a unifying framework for explanation based on the concept of 'understanding.' Understanding, in turn,

[9] Our use of the term follows Thórisson et al. (2019; 2018). The term has appeared elsewhere (cf. Chen and Liu (2016); Fei et al. (2016); Baldassare et al. (2009)) with some overlap in definition.

is dependent on the identification of useful relations (causal and otherwise) that enable, for any given phenomenon Φ, a learner's ability to:

1. Predict Φ
2. Achieve goals with respect to Φ
3. Explain Φ
4. (Re)create Φ

In this approach, explanation is an abstract representation of relationships between an explanandum (that which is to be explained) and a reasonable list of causal chains that, if they were to be changed, would change or remove the explanandum, to the extent of damaging performance on one or more of the above evaluation criteria. Thórisson et al. (2020; 2019; 2018) describe a framework for autonomous cumulative modeling that allows an artificial agent to generate causal networks automatically, through experience, that can be evaluated along the above dimensions, and subsequently use them in its further learning. The approach has been tested on complex tasks involving human-robot interaction (Thórisson et al. (2014)) and as of yet, is the only known approach demonstrating autonomous general learning along the lines discussed here.

The ExH concens the third item in this list, stating that internal processes for explanation generation is crucial for further (self-supervised) learning. A key reason why explanation is relevant to learning, according to the ExH, are the following background assumptions concerning the context in which general machine intelligence may find itself, and which result from the definitions in section 2 above (p. 7-11):

1. Complex task-environment (vast number of elements and relations organized and related at many levels of detail)
2. Severely limited accessibility to the world during any learning period (hidden states)
3. General and autonomous learning (knowledge acquisition and its transfer to other contexts progresses without outside help)
4. Limited memory of a learner (mind can only hold a fraction of world's elements and relations)

A complex environment offering limited accessibility means that information comes in bits and pieces, so learning must be continuous or *cumulative*. Thórisson et al. (2019) describe cumulative learning as a process of *knowledge unification*, whereby new information – whether in agreement with already-acquired knowledge or not – enters by default into a process of being unified with it. In this process, outdated knowledge may be deleted, prior knowledge deemed incorrect replaced, and missing knowledge added. To count as 'cumulative,' the unification must happen frequently relative to the learner's lifetime. For the knowledge thus created to be useful in other circumstances, generalization must also be part of the process, which rests on the ability to identify the relationships between relevant pieces of knowledge (for instance, features of a particular outdoor sport may be useful for a similar but new indoor sport), which in turn requires models of causes and effects.

4. The 'Explanation Hypothesis' in a Nutshell

Given a particular target state, fact, or situation (i.e. measurement), a valid explanation for it may be generated through a process of logical argument, resting on abduction, whereby seemingly relevant competing models of causal (and other) relations are compared and contrasted to identify a proposed temporal sequence, with relevant given conditions, that – if absent – would have lead to a different outcome.[10] With that in mind, the ‘explanation hypothesis’ (ExH) states that:

Explanation generation is a fundamental and necessary process for general self-supervised learning.

To be clear, it is not the *only* process that is involved in general self-supervised learning. It is, however, our main focus in this paper. The foundation of the claim rests on several interlinked assumptions, that may be explicated in the following way:

§1 Autonomous general learning (AGL) – i.e. general self-supervised learning – involves the *creation* of knowledge structures about phenomena *unfamiliar* to a learner. The AGL process, in this view, is a change from knowing very little (or “almost nothing”) about the phenomena at hand, to knowing more about them (all the way to “almost everything”). This is a process of information transformation (actions, measurements, and systematic construction of structured information), facilitated by machinery that we assume exists in the learner before the learning starts; the knowledge thus generated by the learner – without outside teaching assistance of any kind – is produced through an interaction between the learner’s learning mechanisms and its environment, via its body.

§2 A capacity for AGL means that the kinds of spatio-temporal or cognitive patterns an agent can learn (i.e. that may be successfully handled by the agent’s learning machinery) is not overly targeted or limited to particular domains or tasks, but rather, may be of many types and forms. By ‘type’ here is not only meant a re-arrangement of enumerable element classes with enumerable relationship types, but also that new kinds of relationships and elements – unforeseen by the agent’s designers – may be constructed (invented) and subsequently explored and learned about from experience by the agent. This forms a kind of knowledge bootstrapping which assumes the existence of spatio-temporal regularity exhibiting non-random correlation, the ability of an agent to reliably measure the resulting patterns, and having machinery for generating and manipulating models of them.

§3 The knowledge thus created cannot be assumed to be perfect on first try—indeed, it must be assumed that some percentage – even a large proportion – of newly created knowledge is incorrect (useless, or largely so). Therefore, knowledge created this way is defeasible (cf. Pollock (2010)) and must allow revision. For a novel phenomenon (Φ_{nov}) we refer to any proposed parts (dissections) of the phenomenon $\{\varphi_1 \ldots \varphi_n \in \Phi_{nov}\}$, and their relations to each other and other phenomena ($\Re_\Phi$), based (in part or in whole) on a learner’s experience of these, as *hypotheses*.

[10] Or, in the case where the outcome may not be different, then a different temporal sequence and conditions than the one given would stand instead as a valid explanation.

Whether the hypotheses are generated by analogy, random exploration, simple reasoning or some other mechanisms, their generation would be as informed as possible, based on prior knowledge, to avoid wasting time on fruitless models.

§4 Generation of hypotheses about a novel phenomenon Φ_{nov} by a learner must be created in light of its existing knowledge $\mathcal{K}$, that is, any prior knowledge $k \subset \mathcal{K}$ deemed most relevant at some point in time, $Rel_{t1}(\mathcal{K}) = k \mid \mathcal{M}(\Phi_{nov})_{t1} + \mathcal{G}^{t1}_{act}$, where $\mathcal{M}_{t1}$ are the measurements of Φ_{nov} at time $t1$ that will be used to facilitate the hypothesis generation in light of what is known, and $\mathcal{G}^{t1}_{act}$ is the learner's active goal hierarchy at time $t1$ (e.g. to explore Φ_{nov}, to not get hurt, etc.). Out of the (potentially very large and diverse) body of knowledge that a general learner may already possess,[11] a-priori knowledge about which of it may be relevant to *any* new phenomenon cannot possibly be produced beforehand, and k must thus be computed on demand. The creation of hypotheses thus depends (in part) on existing knowledge and current measurements of the novelty, $\mathcal{M}(\Phi_{nov})_{t1}$, and could also not be produced a-priori. The reasons for why a learning process may find some of the prior knowledge more relevant than some other knowledge must in part depend on computed argumentation related to relevant evidence for *why* this must be (or likely is) the case.[12] Such processes must therefore necessarily rely (in part) on abduction, because this is what we call a process of this kind.[13]

The conclusion is that hypothesis generation – that is, the process of figuring out new phenomena – must involve abduction, which is another way of saying that a process of explanation must be involved.

§5 For any learning agent trying to learn novel phenomena in an infinite world, conflicting hypotheses about the phenomena in question, and the relations between their parts, is unavoidable during some part of the learning process. In other words, any agent learning something new will have, for short or long periods, incomplete and/or incorrect knowledge about that subject matter during its lifetime. The learning process is only successful (on average) if conflicts in the agent's knowledge are reduced over time (on average); one target of subsequent new learning, therefore, involves reducing incompleteness and incorrectness. Stated differently, self-supervised reduction of conflict among the set of models of experience – and which constitute the bulk of a learning agent's growing knowledge – is a persistent goal in AGL.

§6 In this conflict resolution process, the usefulness of candidate representations of the phenomena (that is, of discrete, semantically relevant subsets of knowledge), for the various goals that a learning agent may have, must inevitably be evaluated. Such

[11] This means that when an agent knows very little, knowledge creation is slow and limited. We see this naïve state of a general self-supervised learner as a special case of the framework outlined here; for some ideas on how it may be addressed see e.g. Thórisson (2020).

[12] It does not matter whether the evidence on which such argumentation rests is implicit or explicit—the claim here, so far, is that they must exist in some form. The same holds for the process of argumentation itself.

[13] The generation of hypotheses and the estimation of their relative value, as used here, matches directly the two most common definitions of abduction ("inference to the best explanation" on the one hand (cf. Lipton (1991)) and abduction for generating hypotheses on the other (Peirce (1931–1958)).

evaluation involves the use of arguments for and against each hypothesis, and sets of them. These arguments must be logical, that is, they must follow from the results of experience (explicit and implicit "experiments"), and relations, rules, and elements deemed useful for such evaluation. We call a coherent set of such arguments an *explanation.*

§7 According to **§3** and **§4**, for each hypothesis covering to some extent the same phenomena, in whole or in part, an argument and explanation can be generated for why and how it may be more or less useful than other alternative hypotheses.

According to **§5** and **§6**, in the process of reducing incompleteness and incorrectness, arguments must be created for the usefulness (validity) of conflicting hypotheses. This also requires explanation through abduction.

§8 From **§1–§7** above we may draw the conclusion that the process of general self-supervised learning depends on explanation: *generating explanations is an unavoidable necessity in general self-supervised learning.*

The conclusion is that in fact at least two processes requiring abduction/explanation can be identified in general self-supervised learning: One is involved in the generation of hypotheses about novel phenomena, and another is involved in the process of comparing and contrasting competing hypotheses, when incomplete and conflicting knowledge is to be addressed. These processes are likely to be intertwined in natural cognitive systems; in artificial ones there exists the possibility to implement them in a more discrete fashion.

In addition to these two, there is potentially a third role for explanation that we can also see. Two seemingly good explanations may nevertheless not be *equally* good, and two methods for producing explanations may also not be created equal (or equally relevant in particular contexts, or equally context dependent). In this view, the simple explanations produced by a comparison process are themselves hypotheses—about which explanation methods are more useful. To handle this situation, a recursive application of (meta-)abduction may thus be applied to remove inconsistencies produced by inferior methods for comparison of hypotheses.[14]

5. Some Implications of the 'Explanation Hypothesis'

To summarize the arguments for the 'explanation hypothesis' (ExH) above:

1. From **§1–§6** above it follows that autonomous (self-supervised) general learning processes, in the face of novelty, will produce fragmented (partial) information (loosely-connected relational information graphs) that can be improved[15] in light of the implicit and explicit goals and purposes of an agent. This partial information will be comprised of both loosely- and strongly-coupled fragments of information about the agent's experience.

[14] This use of 'abduction' is close to how the term was used originally by Peirce (1931–1958).

[15] We do not say that the knowledge contains "errors" because the right measure for the value of knowledge - i.e. these information graphs - is their usefulness to an agent, not whether they are "correct" or "true," and the "truth" may be, in many (or most) cases, unobtainable.

2. If we define knowledge harmony as the inverse of the number of inconsistencies in the knowledge base as a whole, a key goal of such knowledge creation must thus be to increase harmony in the knowledge set, which unavoidably includes *reducing erroneous knowledge* and *filling in missing information.*

3. For this process, information fragments must be *evaluated* to determine which information is to be kept, improved, and which is to be discarded. This process involves comparisons between information fragments, resting on arguments for these choices.

4. This is as close as one can get to the very definition of abduction: A process for producing explicit, localized arguments for the coherence of a set of statements. Such evaluation must be based on some form of explanation.

This conclusion has some interesting implications that we will now briefly discuss. One implication is that

§9 *general autonomous learning cannot be done without reasoning processes.*

Here we don't mean only the kind of reasoning that we humans may experience when consciously weighing alternatives or options (although this is included), but rather *any* functional processes that implement operations wherein acquired evidence about something is evaluated, and logical arguments generated, to produce an output whose value to a goal-oriented process can be put to the test. In other words: the use of argumentation to evaluate a model (whether represented as a set of statements, a graph, or in some other form) whose usefulness (validity) for particular purposes (goal-oriented behavior) can be measured and tested. The emphasis here is thus on the function(s) that explanation and reasoning provides, rather than its particular implementation and surface characteristics.

With limited sampling, as is inevitable given limited time and energy, a learner seldom has sufficient evidence through direct perception to conclusively settle all relevant details of a plan, action, or situation—doing experiments, or collecting sufficient data through happenstance, would in most cases take too much time. (For instance, we normally don't do rigorous lab experiments to be certain that the cup we are about to drink from is guaranteed to have coffee in it.[16]) But given relevant knowledge from other situations, the particulars of most situations may still be settled pragmatically by bringing in constraints and generalized rules (statements about relations between phenomena) from other experiences (the cup was full and doesn't leak; I only took one sip; a sip is less than a full cup; hence it still holds coffee). When the need comes to answering questions about details, reasoning may thus be employed. This means that to model any subset of a complex infinite-variety world calls for hierarchical models where logic is applied to evaluate the usefulness of the knowledge

[16] This is of course not to say that we do such reasoning tasks all the time while moving around in our everyday life, and neither that our mind does such reasoning tasks subliminally all the time either, as that would be unnecessary and wasteful. We envision such "common sense" operations to be performed with deeply entrenched ("compiled") high-level knowledge that, over many years of learning, has become extremely fast to call upon. A key feature of such deeply trained knowledge in humans is that its composition can still be consciously inspected – i.e. dissected and inspected – using explicit reasoning processes.

in "holistic chunks," incrementally, according to present (often unforeseen) needs. Doing so without some sort of logic is rather inconceivable; the systematic application of logic is, in turn, reasoning.

Given that a majority of theories about explanation involve some form of causal relations, another implication of the ExH is that

§10 *knowledge of cause and effect is fundamental to general autonomous learning.*

Past research has argued this point before (cf. Pearl (2001)), and while this notion is not new, significant research efforts have been spent on purely statistical methods in AI. By definition, cause is directional because a cause cannot appear after its purported effect; ignoring direction in cause-effect relationships and leaving only their statistical properties removes information that is key to any intelligent agent for getting things done (cf. Halpern and Pearl (2005a), Thórisson and Talbot (2018)). For those in AI already working on causal relations learning, the ExH is a 'inverse' argument for something they have already accepted; for others, the ExH is an argument for why statistics alone falls short of delivering general autonomous learning, and by extension general machine intelligence.

Yet another interesting implication of the ExH has to do with reification. Why is this important or relevant here? It is in fact both: To create a new model of a new phenomenon, alternatives must be considered; these alternatives rest in part on particular groupings of what is perceived about the phenomenon (e.g. the changing curve of a cup's handle due to the cup's rotation relative to the viewer) and based on those groupings, implications can be drawn (e.g. that a formerly perceived shape can be achieved by rotating the cup back). The perceptions of the cup's handle are reified into a group (family) of objects with common functionality which may then be given a name ("handle"). For the ExH to hold, a process for reification must be present in a general learning system; reification in turn requires reflection, meaning that some amount of

§11 *reflective capabilities must be prevalent in a general learning system,*

because without it, alternative hypotheses about new phenomena, both having custom features and scope, could not be effectively evaluated against each other on-demand. Human reasoning relies heavily on this when communication revolves around goal coordination (such as where to go out for dinner or what kind of engine a particular aircraft should be outfitted with). Note, however, that reflection is not an all-or-none property: it may be implemented and realized in various ways, to various extents, depending on its mechanisms as implemented in particular cases.

A fourth interesting implication of the ExH that we could look at here is that

§12 *general autonomous learning requires a process of composite model creation.*

This conclusion follows from the above because a complex world like the physical one involves infinite compositionality, yet displays regularity (since otherwise no learning would be possible) at multiple levels of detail. Efficient modeling[17] of that regularity requires hierarchical representation: flat representations (e.g. straight enumeration via if-then statements) in infinite worlds would not simply be grossly inefficient but also, it cannot be generalized

[17] For a definition of 'model' as used in this paper, see footnote 6 on page 10.

in any obvious way to deal with novelty. If there is anything that an infinite world has an infinite supply of, it's novelty. The knowledge of a learner, in this view, consists of a myriad of peewee interconnected models, whose parts can be handled (in context) through reasoning operations.

6. The 'Explanation Hypothesis:' Arguments For & Against

The Explanation Hypothesis rests on several assumptions and definitions, some of which might challenge it if they are found to be incorrect or have faults. Firstly, is it obvious that 'a coherent set of arguments' should count as "explanation"? Secondly, whether we call it 'explanation' or something else, why should knowledge generation need to rely on arguments? And thirdly, why would the application of logic imply reasoning processes? After all, don't deep neural networks (DNNs) apply logic without any reasoning? Or alternatively, don't ANNs implement reasoning in a different way than traditional reasoning engines? A proper answer to these, and other related questions, is at the heart the claim's foundation. A thorough treatment of all relevant issues would require a book; let's look at some of the most obvious ones in the next few paragraphs. We start with definitions, labeled **D1** to **D4**, and then we move on to more involved claims, labeled **C1** to **C3**.

D1 *"The notions of 'argument' and/or 'explanation' that the ExH rests on, are incorrect."*

A common way to understand the concept of 'argument' has been to build upon notions of truth and absolute validity.[18] This requires then some explication of what is meant by these axiomatic Platonic concepts, which transfer the burden of proof to the concept of truth without getting any closer to producing candidate processes that systems in a complex dynamic worlds might employ for learning. Since the physical world is non-axiomatic (we can never be certain that we know everything about how it works), a pragmatic approach must in any case rest on other foundations. For this we can build on the concept of 'usefulness,' which allows us to propose a notion of *knowledge set coherence* that may be subjected to empirical evaluation: Given a set of generalization statements about the relations between particular phenomena a world (which we assume contains regularity), e.g. that flamingos are pink because they eat carotene-rich plants, the value of the statements for getting things done in the world may be certified by comparing them to experiences of the world (e.g. changing the flamingos' diet to change their color). In this approach, given a set of goals to be achieved (whether survival or doing the dishes), statements that are less often correct will be less useful than statements that are more often correct. If those statements are based on models of the world (which they must be, because compact statements of this kind are a product of information structures, which must be manipulated to produce statements), one could say that the arguments that are more often correct are evaluated to be that because the underlying model for producing them is better (for getting stuff done) than the alternatives (or, for those who prefer, more "true"). A knowledge set A is more coherent than a knowledge set B if it produces more often generalization statements that are deemed

[18] See for instance https://iep.utm.edu/argument/ (accessed on Dec. 8, 2021).

correct (according to some roughly generalized measure) and that are *less frequently in contradiction with each other.*

If we consider such generalization statements to be *arguments* for how the world works, then we have captured the meaning of the concept of "argument" as used here. Therefore, the definition of 'argument' as used here is appropriate for our purposes. By extension, 'explanation' can then be cast as a coherent set of arguments that provide plausible evidence for something to be the case, rather than something else (that is in some way mutually exclusive). It is defeasible because new evidence may come to light, and the evidence may be incorrect or inappropriate, yet in light of *available* evidence, it can be argued to be a good (or even best) explanation.

D2 *"The definition of 'knowledge' that the ExH rests on is incorrect / inappropriate."*

Our notion of knowledge is admittedly based on an information-centric view of cognition and not so much on the views of behaviorism or biology (neurology). However, it does not deviate from how this term is used in cognitive science and (increasingly) in neuropsychology, and is in any case compatible with the use of this term in the artificial intelligence literature (cf. Newell (1982)). This line of reasoning seems quite a bit of a long shot.

D3 *"The definition of 'learning' that the ExH rests on is incorrect / inappropriate."*

Our view of learning is rather general, subsuming most if not all notions of learning in use in current AI and psychology research, and would probably be considered to be a textbook definition by most accounts, as the following dictionary definition of learning shows:[19]

1: *the activity or process of gaining knowledge or skill by studying, practicing, being taught, or experiencing something*
2: *knowledge or skill acquired by instruction or study*
3: *modification of a behavioral tendency by experience (such as exposure to conditioning)*

With reference to the definitions detailed in section 2 (pages 7-11), learning can also be defined as a learner's achievement of an (explicit or implicit) goal of getting better (by some chosen metrics) on a task (or family of tasks), in light of a set of goals relevant to that task, over time.

D4 *"The definition of 'reasoning' that the ExH rests on is incorrect / inappropriate."*

We consider here all forms of reasoning (deduction, induction, abduction, as well as analogy), with a special focus on abduction, as this form of reasoning plays an important part in "deconstructing" phenomena and chains of events and has generally been considered the mode of reasoning for producing explanations. With respect to mainstream philosophy, our notions for these four kinds of reasoning are perhaps closer to those of Peirce (1931–1958)

[19] Mirriam-Webster's online dictionary; https://www.merriam-webster.com/dictionary/learning – accessed on October 3rd, 2021.

than some modern or more strictly axiomatic definitions (cf. Mayer and Pirri (1996)). Either way, the arguments behind the ExH do not rely in any way on alternative or unusual definitions of any of these terms.

Note that even though it may be easier to discuss these reasoning processes in light of human cognition, we consider these reasoning processes to be automatable, and thus implementable in software. In our view, the systematic application of logic over an information set containing models of experience, for a variety of purposes, can be implemented in may ways, and can probably be found in a variety of forms in various biological systems, including non-cognitive ones. We come back to this point again briefly in the example of the rabbit in section **C3** on page 22.

Now that we have laid some criticisms against our key definitions to rest, let's turn to arguments involving more complex ideas.

C1 *"The proposed approach for achieving autonomous general learning (AGL) is formulated in a reasoning-dependent way; AGL could or might be achieved through other, and different, non-reasoning methods."*

A two-part reply to this argument is as follows. Because the mechanisms in autonomous general learning (AGL) deal with novelty, information structures that don't exists in the agent's knowledge must be created from scratch (based on current and prior experience). The world is not experienced all at once (and could not be), so information comes incrementally, and thus AGL learning must be *cumulative*. This means that the learning *context* is part of the learning process (e.g. the time and place that the learning takes place must be qualified), so that later use of that knowledge (revisions, updates, comparisons, etc.) can make generalizations about the information perceived and produced in various places at various times under various conditions. To achieve this, the knowledge created must have discernible parts, enabling localized operations on this information. In other words, hypotheses about novel phenomena and their parts must to some extent be discrete (or discretizable), that is, individually addressable for later assessment, comparison, and other knowledge-based operations in light of other information. Whichever way in which these processes for achieving local operations on subsets of acquired knowledge are implemented, they cannot escape the functional equivalent of a reasoning process, more specifically, a process of judgement that produces the results described, that is, a priority list of the available options for each revision of the knowledge.

The evaluation process we described in **§6** bears some features of reasoning, yet it is not formulated specifically in any reasoning pseudo-language but rather, along the lines of a specification of engineering requirements, in light of the desired behavior of the system as a whole. Further, the processes described are a form of explanation because they involve judgement of the usefulness (justification) of particular local information, that is, the relationship between two or more parts of the knowledge, in light of a potentially large set of information structures covering multiple levels of information detail (and resting at least in part on empirical evidence). The process must involve several local information structures (at least two hypotheses, each of which relates at least two or more things, one of which is shared between them) and one piece of evidence (at the very least, but typically more) that favors one hypothesis over the other, resulting in a *judgement*. The judgement of the quality (relevance, usefulness) achieved in this way is essentially one definition of abduction—a form of reasoning. Another indication that this must rely on a kind of reasoning is this: The evidence for making the judgement is chosen not from a predefined set of evidences but rather according to relevance, from a set of (on-demand) computed alternatives, in light of one or more goals, which itself unavoidably involves reasoning processes.

C2 *"The ExH calls for a 'language of thought' (LoT); while many arguments for a LoT have put forward (cf. Rescorla (2019), Fodor (1975)), these are largely philosophical proposals with limited empirical support."*

This argument contains three parts. Firstly, that the ExH calls for a 'language' (we agree that it may, but possibly not in the way you might think) and secondly, that the idea of a LoT has no reasonable arguments supporting it (it does, in our opinion, but again perhaps somewhat differently than you might think). Because it's called 'language' of thought, a typical assumption people might make is that this language must in some way be similar to *natural* language (this is to some extent true, but probably more loosely than we might think). Let's look closer at these.

We assume that models are compositional, made up of parts that can be inspected and manipulated separately from other parts, much like any decomposable physical object.[20] Beyond this, the ExH sees models as information structures with properties that have no particular leanings towards mathematical, architectural, or some other specific model *form*. Further, this view is framed in the context of a general concept of *control*, and as Conant and Ashby (1970) showed, "every good controller of a system must be a model of that system," that is to say, to control some process successfully, *a model of the system* being controlled (its parts, their relations, and their behaviors—in part and in whole) is *necessarily* required. One basic aspect of control is the act of perception or measurement (to assess the results of actions taken), which inherently embodies the concept of representation: Capturing a value in some form that can be manipulated in lieu of the actual phenomenon that originally caused the measurement results. This is the very definition of *representation*. A more generally-accepted foundation for the ExH than this is hard to imagine (without semantic contortions and serious deviations from commonly used meanings of empirical concepts, including some fundamental ones in physics like time, cause-effect, and energy). Further, a language employs patterns at various levels of detail; at the lowest level we may assume various groupings of the sensory patterns (and cognitive process patterns) mentioned, that then form the basis for concepts, objects, etc. (which in turn can be associated with arbitrarily chosen patterns to be used as labels like words and symbols). So with respect to representational aspects, given this prologue, the ExH is not incompatible with the LoT hypothesis.

We assume that the manipulation of information in a thinking mind must be done systematically. 'Systematicity' implies rules, no matter how complex or convoluted, how they are represented and implemented, and whether they are implicit or explicit. If we imagine these rules taking various forms, one of these might be a kind of language. A requirement for a LoT may say, at one ("strict") extreme, that thought requires *natural language* operations on mental representations, with complex human-language grammar, word equivalents, word categories, etc. (This would seem to imply that no animals except

[20] For this definition of 'model,' see footnote 6 on page 10.

humans are capable of thought.) At its other ("loose") extreme, the idea of a LoT simply states that mental operations follow "language-*like*" rules, without giving any specifics about how this might be implemented.[21] In this loose version it is not much different from saying that a 'LoT is some kind of machine that operates on information according to (unspecified) rules.' This, however, is the only way in which a LoT theory might be compatible with the ExH as presented here, because the ExH says little about *how* the purported explanation operations should or must be implemented. In other words, only according to the most loose definition of 'language' does ExH assume a LoT.

Any LoT hypothesis is well supported in the very general sense discussed here; Conant and Ashby's (1970) proof that successful control requires models of what is to be controlled, provides a mathematical argument – in accordance to what the essence of these words must mean to fully support the implied meaning in their most canonical usage – that cumulative modeling (Thórisson and Talbot, 2018) unavoidably results in compositional knowledge; operations on this knowledge must follow some "language" (read: hierarchical rules), as they are used for various purposes in subsequent planning, task execution, as well as when they are changed, extended, deleted, or analyzed for various purposes.

C3 *"The analysis above relies on a top-down, high-level view of cognition, making the invocation of high-level reasoning concepts like abduction easy. Learning paradigms exist that don't rely on concepts inspired by human cognition but rather on biological 'sub-symbolic' approaches, for example animals and artificial neural nets."*

No artificial neural networks (ANNs) – including deep neural nets (DNNs) – learn as of yet cumulatively[22] – and some research has argued they cannot do so by design (cf. Wang and Li (2016)). Their learning happens in one continuous iterative training session, after which their learning capacity is turned off, and thus their learning methods never need to explicitly address discrete subsets of the knowledge they hold at runtime. When we ask that they be explainable, this limitation hits like a sledge hammer: Without the ability to address discrete, salient subsets of knowledge, generating local explanations and evaluations (see **§6** on page 14) is prevented by design. ANNs of all kinds certainly apply logic (via weight functions), but their reasoning is limited at best,[23] and they don't explain anything to themselves or others because to do so they would need the necessary representational foundation (in particular, hypothesizing causal relations of localized, reified knowledge sets—see **§11** on page 17) and machinery (abduction, but also more generally, ampliative reasoning).

This means that local explanations, of the kind argued for here, are not obviously implemented in modern-day ANNs and is part of the reason why they must be trained all-at-once. Some future version of ANNs might be capable of doing cumulative learning—but if the ExH is correct then this would require additions and (significant) modifications, in accordance with the principles outlined in this paper.

[21] We ignore, for convenience and brevity, other implications of the a literal interpretation of the LoT hypothesis, such that e.g. there should be only one thread of processing (since people are largely incapable of creating two linguistic sentences simultaneously), because we find such a reading of it not nearly as productive or useful for research in self-supervised learning as considering it a loose analogy.

[22] See definition of 'cumulative learning' on page 11.

[23] Several authors have argued that the kind of function approximation that ANNs do does not meet the most common definitions of 'reasoning' (cf. Wang (2006), Wang and Li (2016)).

A key feature that separates humans from other animals is an extensive capacity for language. The ability to learn and use language relies on explicit separation of a 'message' from the message 'carrying mechanisms,' and the 'context' in which this is done, depending in large part on inference and systematic handling of a network of inter-dependencies. Much of what enables humans to design complex machines, explain novel phenomena and invent new ideas, seems to rely on abilities that similarly allow isolating goals, constraints, methods, and strategies, and treat them separately, as needed, piecing them together into webs with complex dependencies, in accordance with what the situation may call for. The systematic application of logic seems thus intricately at play not only in our ability to manage linguistic structures but also in our ability to create and follow complex plans, grasp new concepts, and invent abstract rules.[24] The upshot is that the levels of reasoning sophistication, as expressed in the ability to match cognition to environmental and task complexity, may be more of a continuum than a discrete space.

In what sense then does a rabbit rely on (internal) explanations as it goes through its day, avoiding danger, plotting a path through the forest? Is its learning in this case truly explanation-based, as the ExH states? The thesis explored here is that autonomous general learning – insofar that it is *general* (see definition p. 10) – must rely in some ways on abductive explanation. These explanation processes do not need to be explicitly accessible to the learner's cognitive mechanisms for other purposes, e.g. for producing explicit explanations of thought and behavior like humans do (after all, rabbits are not known for explaining anything). In other words, the explanation-generation processess themselves do not need to be reifiable (see **§11**). The explanation form best known to humans is the kind of conscious, thinking-out-loud that we deliberately practice sometimes. To avoid pars pro toto, confusing this with the larger concept of what constitutes an explanation according to the ExH, Thórisson (2020) proposed to use 'micro-ampliative reasoning' for various forms and mixtures of deduction, abduction, induction, and analogy that uses these in more limited (and possibly less distinguishable) ways, and is not based on explicit symbol manipulation like is enabled through the use of (natural) language. The explanation-generation processes must operate on local, reified information (as explained in **§4**, **§6**, **§7**, and **§11**), but there may be different ways and levels to which such requirements are met in a particular cognitive architecture; the cognitive capacities expressed by the architecture will in turn be affected.

Therefore, the answer to this question is *"Yes: Animals (other than humans) are also subject to the ExH."* However, their reasoning mechanisms are probably not implemented in the same way as ours, depending on the particulars of their cognitive architecture (working memory, representational scheme, etc.), and thus their reasoning is subject to somewhat different constraints and properties (cf. Mannella et al. (2021)). As a result, ability to transfer skills across various situations, domains, conditions, and tasks (which is one way to define generality) will vary. A key aspect to keep in mind here is the complexity of the task-environment and a learner's limited memory capacity: For any convoluted multitask-environment, a vast number of inter-dependencies between the various constituents of the world (consider e.g. the giant number of view-dependent images that may fall on the image sensor of a submarine robot's camera during a short underwater trip) make it impractical

[24] Whether language rests on special cognitive structures or shares core learning mechanisms with other skills is surely still an open question, but there can be little doubt that language is not necessary for many skills that humans are capable of and separate them from other animals.

to enumerate these (or provide a finite list of them up-front), due to the unavoidable combinatorial explosion. The obvious solution is to extract layered regularities and use these in a regimented rule-based and combinatorial way, a process requiring explanation à la the ExH.

Note that in this sense, the ExH has limited scope: It simply states that explanation cannot be avoided in autonomous general learning (AGL). The role of abductive explanation in AGL is to handle variety in a learner's experience[25] that would otherwise render a cognitive apparatus incapable of practical learning. Depending on various other related factors, including how general and autonomous the associated learning mechanisms are, its runtime characteristics may thus be expressed in a variety of forms, including animals that cannot explain anything to others, yet rely on such mechanisms for learning.

7. Conclusions

We have presented a compact version of what we call the 'explanation hypothesis' (ExH)—the claim that general self-supervised learning necessarily requires processes of explanation. The argument, in short, rests on the premise that general learning – which, to some extent, is independent of what is being learned – can only proceed incrementally and cumulatively, and that errors are unavoidable during this process. So a general learner needs to keep track of the why and the how of its own knowledge acquisition, with accompanying arguments for why certain learning approaches lead to better-quality knowledge than others. A process of this kind must unavoidably rest on argumentation, which in turn requires abduction (what may have caused what, what leads to what, etc.). Abduction is otherwise known as explanation.

Based on fairly standard (yet slightly more specific) definitions of the key concepts involved, we have looked at several arguments for and against this claim, as well as a detailed argument for how it may be construed. The implications of the ExH, should it turn out to be valid, are many and varied. Perhaps the most significant implication is that a general learner must be capable of reasoning. While this may not be news to some AI researchers, the most prevalent paradigm in the field for the past decade has been artificial neural nets (ANNs); it is not known how – or if – these could be outfitted with a reasoning mechanism of the kind outlined here. Another deep implication is the need for compositional knowledge representation, and the need for *semantically localized* operations on that knowledge, both implying fundamental mechanisms for on-demand *knowledge reification*.

The arguments we have considered against the claim have come from various angles. Although we have most certainly not looked at an exhaustive list of arguments that could potentially be fielded against the ExH, the more obvious ones we looked at here seem to be rather effortlessly deflected. Neither our definitions on which the outline here rests, nor the larger-scope ones referencing other AI approaches, seem to penetrate very much its basic stance.

[25] By 'handling variety' is meant the tracing of sources of differences (e.g. the particular outcomes of using two or more strategies for achieving particular kinds of goals in particular kinds of circumstances) to its 'root causes' (or simply: useful knowledge of cause-effect relationships), so that general rules may be derived, e.g. through induction or analogy.

Future research will have to be undertaken to explore the empirical foundations of this claim. We consider it likely that research on non-axiomatic logic and reasoning might be useful for further formalizing the ideas put forth in this paper. For this purpose, cognitive architectures demonstrating some domain independence, cumulative learning, and introspective capabilities, are needed; promising candidate frameworks include NARS (Wang (2013)), Leela (Kommrusch et al. (2020)), and our own AERA (Nivel et al. (2013)). In the mean time, the explanation hypothesis stands open both inspire new research into general machine intelligence and to further scrutiny.

Acknowledgments

I would like to thank my long-time collaborators Eric Nivel and Pei Wang for numerous insights and discussions about the topics addressed in this paper, as well Jeff Thompson and my students Arash Sheikhlar, Leonard Eberding, and Hjörleifur Henriksson. Many thanks also to Paul Robertson and Steve Kommrusch who provided useful comments on an earlier manuscript. This work was supported in part by the Department of Computer Science at Reykjavik University, the Icelandic Institute for Intelligent Machines, and a grant from Cisco Systems Inc. —*KRTh*

References

Gianluca Baldassare, Marco Mirolli, Francesco Mannella, Daniele Caligiore, Elisabetta Visalberghi, Francesco Natale, Valentina Truppa, Gloria Sabbatini, Eugenio Guglielmelli, Flavio Keller, and others. The IM-CLeVeR project: Intrinsically motivated cumulative learning versatile robots. In *9th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 189–190, 2009.

Jordi Bieger and Kristinn R. Thórisson. Evaluating understanding. In *IJCAI Workshop on Evaluating General-Purpose AI, Melbourne, Australia*, 2017.

Zhiyuan Chen and Bing Liu. *Lifelong Machine Learning.* Morgan & Claypool Publishers, 2016. ISBN 1627055010, 9781627055017.

Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

Geli Fei, Shuai Wang, and Bing Liu. Learning cumulatively to become more knowledgeable. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1565–1574, 2016. ISBN 978-1-4503-4232-2.

Jerry Fodor. *The Language Of Thought.* New York: Thomas Y. Crowell, 1975.

Óscar Fontenla-Romero, Bertha Guijarro-Berdiñas, David Martinez-Rego, Beatriz Pérez-Sánchez, and Diego Peteiro-Barral. Online machine learning. *Efficiency and Scalability Methods for Computational Intellect*, pages 27–54, 2013.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach — part i: Causes. *Brit. J. Phil. Sci.*, 56:889–911, 2005a.

Joseph Y. Halpern and Judea Pearl. Causes and explanations: A structural-model approach — Part II: Explanations. *Brit. J. Phil. Sci.*, 56:843–847, 2005b.

Steve Kommrusch, Henry Minsky, Milan Minsky, and Cyrus Shaoul. Self-supervised learning for multi-goal grid world: Comparing leela and deep q network. In *Proceedings of Machine Learning Research*, volume 131, pages 81–97, 2020.

Peter Lipton. *Inference To The Best Explanation.* New York: Routledge, 1991.

Francesco Mannella, Federico Maggiorea, Manuel Baltierib, and Giovanni Pezzulo. Active inference through whiskers. *Neural Networks*, 144:428–437, 2021.

Marta Cialdea Mayer and Fiora Pirri. Abduction is not deduction-in-reverse. *Logic Journal of the IGPL*, 4:95–108, 1996.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. *Commun. ACM*, 61(5):103–115, 2018. ISSN 0001-0782.

Allen Newell. The knowledge level. *Artificial Intelligence*, 18:87–127, 1982.

Eric Nivel, Kristinn R Thórisson, Bas Steunebrink, Harris Dindo, Giovanni Pezzulo, Manuel Rodriguez, Carlos Corbato-Hernandez, Dimitri Ognibene, Jörgen Schmidhüber, Ricardo Sanz, Helgi P. Helgason, and Antonio Chella. Bounded recursive self-improvement. *Tech report RUTR-SCS13006, Reykjavik University – School of Computer Science*, 2013.

Judea Pearl. Bayesianism and causality, or, why I am only a half-Bayesian. volume 12, pages 19–36, Corvallis, OR, 2001. Kluwer Academic Publishers.

Charles Sanders Peirce. In *The Collected Papers of Charles Sanders Peirce, 1931–1958.* Harvard University Press, 1931–1958.

John L. Pollock. Defeasible reasoning and degrees of justification. *Argument and Computation*, 1(1):7–22, 2010.

Michael Rescorla. The language of thought hypothesis. In *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, 2019. URL https://plato.stanford.edu/archives/sum2019/entries/language-thought/.

Daniel L. Silver, Qiang Yang, and Lianghao Li. Lifelong Machine Learning Systems: Beyond Learning Algorithms. In *AAAI Spring Symposium: Lifelong Machine Learning*, 2013.

Kristinn R. Thórisson. Seed-programmed autonomous general learning. In *Proceedings of Machine Learning Research*, pages 32–70, 2020.

Kristinn R. Thórisson and Arthur Talbot. Cumulative learning with causal-relational models. In *Proc. Int. Conf. Artificial General Intelligence*, pages 227–237. Springer, 2018.

Kristinn R. Thórisson, Eric Nivel, Bas R. Steunebrink, Helgi Páll Helgason, Giovanni Pezzulo, Ricardo Sanz, Jürgen Schmidhuber, Haris Dindo, Manuel Rodriguez, Antonio Chella, Gudberg K. Jonsson, Dmitri Ognibene, and Carlos Hernandez. Autonomous Acquisition of Situated Natural Communication. *Computer Science & Information Systems*, 9(2):115–131, 2014. Outstanding Paper Award.

Kristinn R. Thórisson, David Kremelberg, Bas R. Steunebrink, and Eric Nivel. About understanding. In *Proceedings of the International Conference on Artificial General Intelligence*, pages 106–117, New York, NY, USA, 2016. Springer-Verlag.

Kristinn R Thórisson, Jordi Bieger, Xiang Li, and Pei Wang. Cumulative learning. In *Proceedings of the 12th International Conference on Artificial General Intelligence*, pages 198–208, Shenzen, China, 2019. Springer.

Pei Wang. *Rigid Flexibility.* Springer, 2006.

Pei Wang. *Non-Axiomatic Logic: A Model of Intelligent Reasoning.* World Scientific Publishing, Singapore, 2013. ISBN 978-981-4440-28-8.

Pei Wang and Seemal Awan. Reasoning in non-axiomatic logic: a case study in medical diagnosis. In *Artificial General Intelligence*, pages 297–302. Springer, 2011.

Pei Wang and Xiang Li. Different conceptions of learning: Function approximation vs. self-organization. In *Proceedings of the International Conference on Artificial General Intelligence*, pages 140–149, 2016.

Yusen Zhan and Matthew E. Taylor. Online Transfer Learning in Reinforcement Learning Domains. *arXiv preprint arXiv:1507.00436*, 2015.

Du Zhang. From one-off machine learning to perpetual learning: A step perspective. *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.