Evaluating Understanding

Jordi Bieger[1] & Kristinn R. Thórisson[1,2]
[1] Center for Analysis and Design of Intelligent Agents,
School of Computer Science, Reykjavik University, Iceland
[2] Icelandic Institute for Intelligent Machines, Iceland
{jordi13,thorisson}@ru.is

Abstract

Understanding is an important aspect of intelligence that has taken a back seat in many approaches to AI. While results in automation can be achieved without it, we argue that understanding is especially important for general-purpose systems. Understanding goes beyond "good performance" on a range of dimensions: if we know that a system understands, we can trust that it will behave relatively robustly, reasonably and predictably—even in novel situations—and that it will be able to use previous understanding to facilitate the acquisition of new understanding. It is doubtful that we could classify systems as having general intelligence if they don't really understand their tasks, environment, and world, and thus it is important for us to verify the level of understanding of any system intended to strive for generality and autonomy. But because understanding is a hard-to-define internal property of a system, evaluation can be difficult. To further our understanding of understanding and facilitate the development of understanding AI systems, we propose four kinds of tests: A system is said to understand a phenomenon if it can make predictions about it, achieve goals with respect to it, explain it and (re)create it.

1 Introduction

Understanding is a core aspect of intelligence. It implies an ability to not only know facts and perform skills by rote memorization ("blindly"), but to be able to reason about them, dissect them, achieve goals with them, and generally *explain* them. Despite its importance, understanding has received relatively little attention in the field of AI, and even in philosophy (cf. [Thórisson *et al.*, 2016b] for a short overview).

In prior work we have characterized understanding as the construction and use of a causal-relational model that incorporates causal as well as other relevant relations of the phenomenon to be understood [Thórisson *et al.*, 2016b]. For some tasks understanding is not strictly necessary; memorization can for instance work fine in tasks that don't change, and knowledge of mere correlations (as opposed to causation) is often sufficient for prediction. But for general-purpose AI or artificial general intelligence (AGI) it is difficult to see how we could get away with a system having no understanding of its task, environment, or purpose.

A major hallmark of AGI is the ability to operate successfully and robustly in a wide range of task-environments[1], even ones that were not necessarily envisioned by the system's creators. Doing this requires more than memorization and learning of statistical patterns. Without an understanding of what made the system's knowledge true in old situations, there is no way to systematically transfer and adapt that knowledge to new situations[2], and the system would break down, behave unpredictably, or need to start learning from scratch. AI systems that initially seemed to perform very well can fail in ways that seem utterly baffling to a human observer, because they are oversensitive to changes, no matter how large or tiny, that were not present during learning/training [Szegedy *et al.*, 2013; Nguyen *et al.*, 2015].

Understanding is not only important for an AI/AGI system itself, to be more capable in its endeavors, but also for our own trust in such systems: There should be some way for us to ascertain that they understand what we entrust them with, that they understand their own limitations, and that they can acquire new understanding when needed. Almost by definition, we cannot test every possible situation such a system could find itself in — and certainly not future circumstances that are unknown at the system's design time. We may, however, want to reliably know how it might adapt to various *kinds* of future circumstances, events and conditions[3], and this is where understanding enters the picture. This paper addresses the question of how we can evaluate whether an AI system does indeed understand.

---

[1] We consider the separation of task from the environment in which it is performed to be somewhat arbitrary, so we use the term task-environment to refer to all aspects of a job that a machine would be doing [Thórisson *et al.*, 2016a].

[2] Much work has been done on "structure mapping" and "predictive analogy" where recognition of similarities between a known (base) domain and new (target) domain allow for the mapping/translation of the base domain's extended understanding onto the target domain [Gentner, 1983; Schmid *et al.*, 2003].

[3] It is important to be able to ascertain that a highly capable system truly understands its (and our) goals [Steunebrink *et al.*, 2016], both to avoid concrete AI safety issues today [Amodei *et al.*, 2016] and hypothesized catastrophes in the longer term [Bostrom, 2014].

Note that this is different from, but related to, the question of *comprehensibility*: how an observer or system designer can understand the behavior and decisions of an AI system [Besold *et al.*, 2016]. The difference lies in who is attempting to do the understanding—we or the AI system—and what is being understood — the AI's decisions or some phenomenon of interest to the AI. A similarity is that we are trying to learn something about internal mental processes of the AI system in order to increase our trust in it. Because of the need to trust AI systems that are increasingly affecting the way we perform important, high-impact, and safety-critical tasks, the field's interest in "explainable AI" (XAI) has grown stronger recently [Gunning, 2016].

In our approach to evaluating understanding, explanation is one of four criteria we should test for: in addition to making good predictions of a phenomenon, a system that truly understands it should also be able to achieve goals with respect to it, explain it, and eventually (re)create it.[4]

In the next section we will describe the various background concepts required for talking about the evaluation of understanding in an AI system. Section 3 will then discuss four aspects of understanding and their evaluation: prediction (Section 3.1), goal achievement (Section 3.2), explanation (Section 3.3) and (re)creation (Section 3.4). We will then briefly reflect on the difference between evaluating understanding of a phenomenon and evaluating the ability to acquire understanding in general in Section 4. Finally, Section 5 has some concluding remarks and future research directions.

2 Background Concepts

2.1 System-World Interaction

For the purposes of this paper, we take a somewhat "dualistic" view of the interaction between an intelligent system and its world[5] [Thórisson *et al.*, 2016a]. Intelligent systems continually receive inputs/observations from their environment and send outputs/actions back. Some of these inputs may receive special treatment — e.g. as feedback or a reward signal.

The world W consists of a set of variables $\mathcal{V}$, dynamics functions $\mathcal{F}$, an initial state $\mathcal{S}_0$, domains $\mathcal{D}$ of possible values and ranges for those variables, possible relations $\mathcal{R}$ between the variables, and a possibly empty set of initially true relations $\mathcal{R}_0$: $\mathcal{W} = \langle \mathcal{V}, \mathcal{F}, \mathcal{S}_0, \mathcal{D}, \mathcal{R}, \mathcal{R}_0 \rangle$. The variables $\mathcal{V} = \{v_1, v_2, \ldots, v_{\|V\|}\}$ represent all the things that may change or hold a particular value in the world. The dynamics can intuitively be thought of as the world's "laws of nature", and are viewed here as an automatically executed function that periodically or continually transforms the world's current state into the next: $\mathcal{S}_{t+\delta} = \mathcal{F}(\mathcal{S}_t)$. The state-values of any element or variable in the task-environment at any point in time is determined by a causal chain, and we can visualize this in a directed acyclic graph (Figure 2 shows the graph corresponding to the environment introduced in Section 2.3).

Each variable v may take on any value from the associated domain $d_v \in \mathcal{D}$. For physical domains we can take the domain of each variable to be a subset of the real numbers. Invariant relations $\mathcal{R}_I$ are Boolean functions over variables that hold true in any state that the system will ever find itself in. In a closed system (with no outside influences) the domains and invariant relations are implicitly fully determined by $\mathcal{F}$ and $\mathcal{S}_0$. In an open system—where change may be caused externally—explicit definition of domains and relations can be used to restrict the range of possible interactions.

Environments are views or perspectives on the world. In their simplest form they can be characterized as slices or subspaces of the world, where the variables are a subset of the world's variables, each variable's domain is a subset of that variable's domain in the world, and only the relevant dynamics and invariants are inherited.

2.2 Models, Phenomena, & Understanding

Our theory of understanding rests on the notion of minds, or agents, making *models* of phenomena external to themselves. These models are either atomic or hierarchically made up of (sub)models that can be used and re-used to model other phenomena [Thórisson *et al.*, 2016b]. A model $\mathcal{M}$ of a phenomenon Φ is denoted M_Φ.

A phenomenon Φ similarly consists of a set of elements $\{\phi_1 \ldots \phi_{\|\Phi\|} \in \Phi\}$, which can consist of other phenomena, variables, and relations $\mathcal{R}_\Phi$ (causal, mereological, etc.). A phenomenon is any grouping of variables and relations in the world that we choose to group as such; $\Phi = \langle \mathcal{V}_\Phi, \mathcal{R}_\Phi | \mathcal{V}_\Phi \subseteq \mathcal{V}_\mathcal{W} \wedge \mathcal{R}_\Phi \subseteq \mathcal{R}_\mathcal{W} \rangle$. $\mathcal{R}_\Phi$ couples elements of Φ with each other, and with those of other phenomena in the world [Thórisson *et al.*, 2016b]. These can be partitioned into inward facing relations $\mathcal{R}_\Phi^{in}$ between element pairs $\phi_i, \phi_j \in \Phi$ and outward facing relations $\mathcal{R}_\Phi^{out}$ between element pairs $\phi_i \in \Phi$ and $\psi_j \in W$. An agent's understanding of a phenomenon Φ is perfect if the model(s) it possesses of Φ *accurately* and *completely* capture Φ's relations $\mathcal{R}_\Phi$.[6] An agent whose models are only accurate for $\mathcal{R}_\Phi^{in}$ understands Φ but not Φ's relation to other phenomena; if models are only accurate for $\mathcal{R}_\Phi^{out}$ it understands Φ's relation to other phenomena but will have limited or no understanding of Φ's internals.

M_Φ is thus a set containing models of a phenomenon Φ $\{m_1 \ldots m_{\|\mathcal{M}_\Phi\|} \in \mathcal{M}_\Phi\}$. The closer the information structures $m_i \in \mathcal{M}_\Phi$ represent elements (sub-parts) $\phi \in \Phi$, at any level of detail, including their couplings $\mathcal{R}_\Phi$, the better a system with models $\mathcal{M}_\Phi$ understands Φ.

[4] By recreate we don't mean *physical* recreation (which would put any part of the solar system out of reach of understanding) but rather to put forth a *necessary and sufficient model* of the *whole* phenomenon such that no aspect of it remains unexplained.

[5] The formulation of an intelligent system (or agent) interacting with the world (or environment) is most commonly used in control theory and reinforcement learning. However, it is a fully general formulation, that also covers traditional cases of e.g. supervised and unsupervised learning. Here the environment simply presents a (training) datum at each time step, the agent responds with a classification or prediction, and—in the case of supervised learning—the environment replies with the target outcome or an error signal.

[6] For any complex phenomenon in a complex world, completeness of $\mathcal{R}_\Phi^{out}$ is generally not to be expected, as this may be an extraordinarily large number. However, for any two phenomena Φ and Ψ that are related, if $\|\mathcal{R}_\Phi^{out} \cap \mathcal{R}_\Psi^{out}\| = small$ then understanding $\mathcal{R}_\Phi^{out}$ may not require a broad understanding of Ψ, even if $\|\mathcal{R}_\Psi^{in}\| = large$.

There is a correspondence between our theory of task-environments [Thórisson *et al.*, 2016a] and our theory of understanding [Thórisson *et al.*, 2016b]. The elements of a phenomenon are directly related to the elements of environments: both have variables, can be hierarchically organized (into sub-environments or sub-phenomena), and contain some form of relations between these elements. As such, we can use much of our ability to reason about task-environments to reason about phenomena for understanding, and vice versa.

2.3 Circuit Example

Throughout the paper we will use the example of understanding about electronic circuits (see Figure 1). An understanding of e.g. AND, OR and NOT gates can occur at the electronics level (top row of Figure 1) and include inward facing relations between subcomponents, or it can occur at the logic level where they have certain abstracted behaviors that can be linked together into larger systems, like an OR gate built out of AND and NOT gates or a full binary adder (bottom row of Figure 1). Below the electronics level lies the chemico-physical level, and above the logic level lies e.g. the context for use in a larger system, for various purposes, etc.

The full binary adder serves as an example of a task-environment to be modeled using variables and causal relations between them (see Figure 2). The adder takes three binary inputs A, B and C, and produces two binary outputs that represent the sum of A, B and C, which can be 0 (00), 1 (01), 2 (10) or 3 (11). To make things more interesting and relate the example more closely to real world tasks, we have added delays to the circuit (this could be achieved using capacitors on the electronic level, or digital circuits—the exact method is immaterial for our purposes). Delays introduce a need for reasoning about time, for instance acting fast enough to achieve goals, predicting temporal behavior of the circuit, or explaining why one must wait for certain events to happen when starting from a particular state.

3 Signs of Understanding

Evaluating understanding in our approach is similar to evaluation of a human/animal's ability to do anything: the more we test, and the more thoroughly we test, the more reliable and trustworthy our estimation of the tested ability will be.

To evaluate understanding of a phenomenon Φ we will look at four performance criteria of a purported understander:

- the ability to make predictions about Φ
- the ability to achieve goals with respect to Φ
- the ability to explain Φ
- the ability to create or recreate Φ

A *thorough* evaluation of understanding of a phenomenon would involve tests from *all four categories above*. A least-thorough evaluation would be testing only the first category; adding subsequent categories lower down will gradually increase the quality of the evaluation. In essence, evaluating understanding involves creating a *map* of *an understander's ability to understand* some phenomenon—the more we sample, and the wider, the more complete and reliable this map will become.

3.1 Prediction

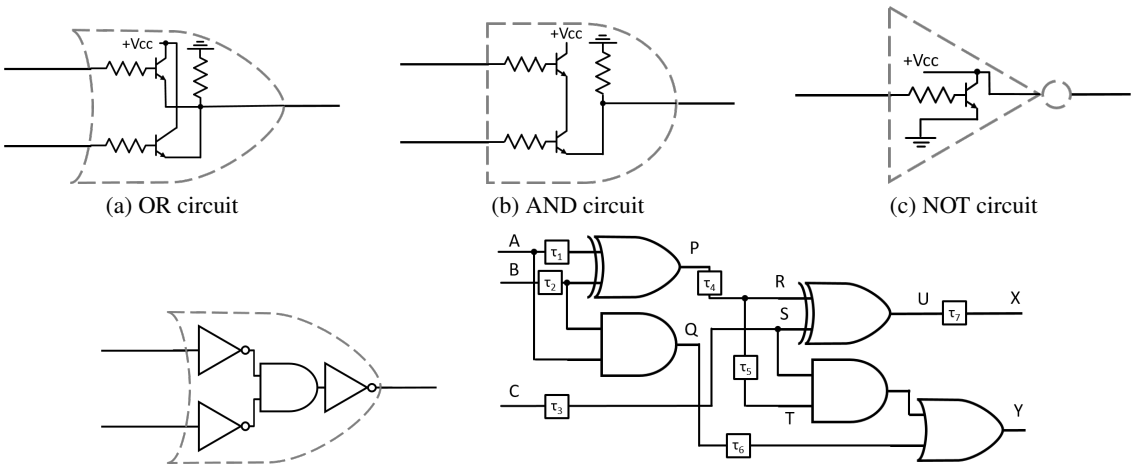
Understanding of a phenomenon should involve knowledge of the structure, patterns, invariants and behaviors of that phenomenon. This suggests that given state information about some of the phenomenon's elements/variables at some point in time, it should be possible to make predictions about possible states of other elements and/or other points in time.

Here we refer to "predictions" in a very general sense. We are not restricting ourselves to predictions that are forward in time or causally downstream from the states we provide information about: if you're provided information that the output of an AND gate was 1, then it is predictable that the inputs were both 1 as well. In scene understanding one could e.g. predict which pixels belong together (segmentation), how a region should be labeled (classification), and what things are going on (annotation), based on information in the image [Li *et al.*, 2009]. But we can go even further with a correct causal model — e.g. predicting that the blob on top of a horse is a human, what the intentions of present actors might be, or what the scene might look like in the near past/future based on a physical movement model [Zelinsky, 2013]. To test text understanding we may see if the AI can predict the correct answer to a question (see e.g. [Prager *et al.*, 2000; Levesque *et al.*, 2012]). Predictions can occur forward, backwards or parallel in time, or even all at once.

Predictive ability can be tested using different kinds of questions. In each of these, the evaluated entity must receive some input information $\mathcal{I}$, which takes the form of a set of tuples of a variable v, a time t, and the state s of that variable v at time t: $I \subset \{\langle v, t, s\rangle\}$. The query set $\mathcal{Q}$ then has the same form, but possibly omits some values. If nothing is omitted, this results in questions of the form "will the queried variables take on the queried values at the queried times?". Omitting the states results in questions of the form "what are possible and likely joint state-value assignments for the set of queried variables at the queried times?". Omitting the times results in questions of the form "at what times (if ever) do you expect the queried variables to take on the queried values?". Omitting the variables results in questions of the form "what variables can we expect to take on the queried values at the queried times?". More complexly structured questions could be constructed by omitting different parts of each tuple in $\mathcal{Q}$, or even just *partially* omitting some values (e.g. to indicate ranges of acceptable values). Sometimes, or often, multiple answers are possible, so the system should respond with an answer set A that is a collection of tuples consisting of a confidence value and an information set of the same form as $\mathcal{Q}$ but with all values filled in.

Constructing tests of predictive ability then involves selecting some input information $\mathcal{I}$ and some query set $\mathcal{Q}$. The variables in $\mathcal{I}$ and $\mathcal{Q}$ must be connected through some causal chain, or else prediction of the missing values in $\mathcal{Q}$ from $\mathcal{I}$ is impossible. At least some of this causal path should overlap with the variables and relations in the phenomenon Φ under test. In fact, a proper test suite would not just randomly select $\mathcal{I}$ and $\mathcal{Q}$ to have this feature, but would attempt to get a comprehensive coverage of V_Φ, $\mathcal{R}_\Phi^{in}$ and $\mathcal{R}_\Phi^{out}$.

It may be tempting to just select every possible input into $\mathcal{I}$ and every possible output into $\mathcal{Q}$, so as to completely cover

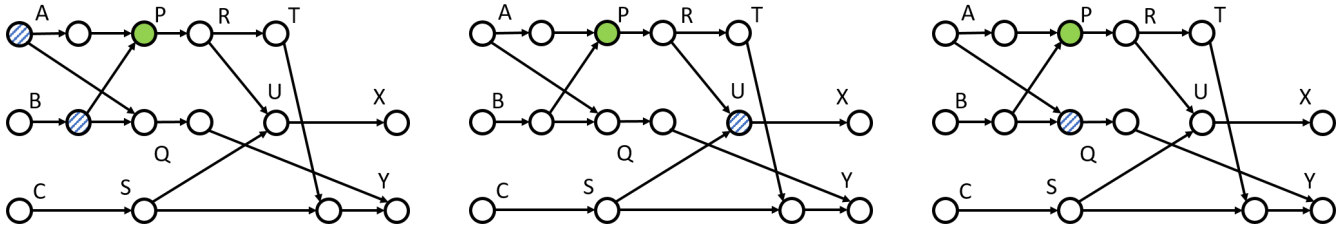


(a) OR circuit (b) AND circuit (c) NOT circuit

(d) OR gate constructed from NOT and AND gates.

(e) Full adder with delays (designated by the boxes with τ_i in them).

Figure 1: Examples of electric circuits. The top row shows the electrical view, while the bottom row shows the logical view.



(a) deduction (b) abduction (c) correlation

Figure 2: Variables corresponding to the adder circuit in Figure 1e including their causal relations, with different prediction tests. The green full circle represents the query variable, while the dashed blue circles represent the input information. Based on their relative positions, different modes of reasoning are necessary.

everything about Φ in a small number of tests. In our example, this means putting variables A, B and C in $\mathcal{I}$ and X and Y in $\mathcal{Q}$ (see Figure 2). However, it would still be necessary to test a good sampling over possible input values for A, B and C, as well as times t for each variable. Furthermore, failure would only suggest that *some* understanding is probably lacking, but not what part. To get a complete map, a more granular and scientific approach must be adopted where understanding of different parts is gradually ascertained.

Another important consideration is how the relative locations of variables+times in $\mathcal{I}$ and $\mathcal{Q}$ affect the kind of reasoning required to produce an answer. For instance, if the inputs are upstream from the query variables, we would be testing deduction, and if they provide full coverage of the causal influences on the query variables, we expect an understanding entity to answer correctly (see Figure 2a). If the inputs are downstream from the query variables, we are testing abduction, which—unlike deduction—cannot typically guarantee the correctness of its inference given correct inputs, even if the system fully understands the related phenomena (see Figure 2b). In this case the inputs represent outcomes, that presumably constrain the possible values of the query variables. In other cases, input and query variables may occur more or less in parallel, in which case the answer can be obtained through (statistical) correlation (see Figure 2c), or through a sequence of abductions (to find a common cause) followed by deductions, or through a sequence of deductions (to find a common affected outcome variable) followed by abductions.

Success on these tests provides some evidence that understanding is present, and failure suggests that *something* went wrong. Wrong answers can be caused by a lack of understanding, but also e.g. by an inability to perform abductive reasoning or output multiple outcomes. Such capabilities should be taken into account in the design of tests, and when failure occurs its source should be uncovered. Some common considerations include selecting only observable variables for $\mathcal{I}$ and $\mathcal{Q}$, and only selecting combinations that sufficiently constrain the correct answer to a single value.

3.2 Goal Achievement

Many definitions of general intelligence refer to a system's ability to achieve complex and diverse goals in a wide variety of situations [Legg and Hutter, 2007]. This paradigm of goal achievement—albeit not always in a general, domain-

independent manner—has been embraced by various AI branches like planning and reinforcement learning. Understanding should almost certainly help in the achievement of complex and diverse goals. To test this, we must assign the system a "task" that involves the phenomenon in some way.

A task in a particular environment consists of an initial starting state (variables in the environment having particular values), a goal (a set of states with desirable joint variable assignments that must be achieved at particular times), optionally come constraints (states that result in instant failure), a body (set of observables and affordances) and some limited information about the task. For instance, in our example circuit, we might start at time $t = 0$ with all variables set to 0, the goal being to set X to value 1 before $t = 10$ under the constraint that Q must always be at 0, with the ability to affect A and B and to observe C, and knowledge of all the delays in the system. An understanding of how delays work would tell the system that if C (or U) holds some value at time t, then S (or X) will hold that same value at $t + \tau_3$ (or $t + \tau_7$), and an understanding of XOR gates will tell the system the value of U at time t given values for R and S at time t.

Goal achievement tends to involve planning, or backward chaining from the goal. An agent with an understanding of both delays and XOR gates should be able to know that in order to set X to 1 at time t, U must be 1 at time $t-\tau_7$, which means that we need $R = 1$ and $S = 0$ or $R = 0$ and $S = 1$ at time $t-\tau_7$. While an AI might not be able to output a detailed plan in a symbolic language, it might be able to demonstrate understanding by doing. A system that could do both would of course be preferable as one would have more options for how to test it. Since S is determined by C at time $t-\tau_7-\tau_3$, it is not under the system's control, and it must respond to what it observes the value of C to be: if it's thought that S will be 0 at some time, A and B should be controlled to produce a 1 at R at that same time, while avoiding setting Q to 1 as well. How the task should be executed depends on the exact values of the delays, and in fact the task may be impossible to perform for an agent if τ_3 is short enough and C is controlled by an adversarial process. A task like this could be used to evaluate understanding of the internals of the entire "adder" circuit $\mathcal{R}^{in}_{adder}$, as well as e.g. the outward relations of XOR gates $\mathcal{R}^{out}_{XOR}$ (see Figure 1e).

As with the construction of prediction tests, constructing goal achievement tests requires selecting different sets of variables, and settings for them. The variables need to be selected in such a way that there is a causal path from the control variables to the goal and constraint variables that overlaps at least partially with the elements of Φ. These variables can be selected randomly, although a good test suite would need to ensure proper coverage of all elements of Φ, and ideally different combinations of elements and outward relations. An initial state can be created by starting from any realistic initial state (or some "null state") and executing some (possibly random) behaviors on selected nodes. To determine (un)desirable state values to serve as a goal (and constraints), another (random) behavior could be executed on the nodes that the AI system will get to control from the initial state and some future state could be designated as the goal. Repeating this behavior would then be one way of solving the task, which ensures that this is in fact doable. The difficulty can be increased by executing multiple behavior sequences from the start state, and setting the goal to be a state that only occurs rarely. To construct a good test—as with the other three methods discussed in this paper—good understanding of the target phenomenon Φ is necessary.

3.3 Explanation

When humans want to evaluate each other's understanding of a phenomenon they often ask the understander to explain it to them, in part or in full. Explanation then involves summarization of the relevant elements of the phenomenon[7], and tracing it back to the most salient causes. This requires a capability that is currently lacking in virtually all AI systems. For this reason, initiatives like DARPA's Explainable AI (XAI) were started [Gunning, 2016]: AI should not just make good decisions—which is of course important—but also be able to explain them so that humans can understand and trust them.

In this paper we are ultimately more interested in whether *the system* understands the causal nature of the phenomena it observes, manages, or controls *for itself*. This means being able to identify the *necessary* and *sufficient* variables and their relations to explain *precisely* what we ask it to explain at a desired level of detail/abstraction — it is not enough here to get a printout of the system's full knowledge base. However, in theory it should not have to cater its answer specifically to humans, as this would not only require understanding of some phenomenon Φ, but also of humans and/or human natural language. An AI system is not necessarily working with human models (our own ground truth) of phenomena, and would need to answer in terms of actions and observations that it itself can make, or in terms of its own models $\mathcal{M}_\Phi$. Still, in practice interpretation of these models is potentially problematic unless the system has some method for presenting output in a reasonably human-readable way.[8]

In our circuit example, some actions will have particular effects. We expect an understanding system to be able to highlight/select variables that are responsible for the effects at a desired level of abstraction and detail; for circuits an explanation could be at the level of half-adders, logic gates, or electronics, depending on what the system is being asked to explain. Furthermore, we want to know what causes are *salient*: if we ask why the output of an AND gate became *true* when input A was always *true*, while input B only recently became *true*, a proper explanation would highlight input B.

Typically, when we're asking for an explanation, we don't want to hear "the output is a 'dog', because pixel (1,1) had value [255,255,0], pixel (1,2) had value [255,240,10], etc." — while this is one form of explanation it is not the kind of necessary-and-sufficient identification of causal relations that we are after. In this example we would want something like "it's a dog because it has I see lots of fur, two ears, two eyes,

[7] cf. [Nenkova and McKeown, 2011] for how automatic summarization is used in language understanding.

[8] Much work is currently done on interpretability of AI systems (see e.g. [Lou *et al.*, 2012; Kim, 2015; Ribeiro *et al.*, 2016; Besold *et al.*, 2016]), and we're hopeful this problem can eventually be overcome.

a nose, a mouth, four legs and a leash". And then we should be able to go deeper: "why do you think that's an ear?"

Moving between levels of detail and abstraction when providing explanations is important as this is one way to uncover the boundaries of the understander's model — whether the answers are e.g. derived from a set of memorized statements or whether the purported understander has a compact causal-relational model, and how far it reaches. Another way to test for boundaries is to introduce novel modifications or inputs to the task-environment, e.g. "I know you know how to play tennis, but how would you play tennis in micro-gravity?" For an understander that doesn't understand either microgravity or tennis this becomes virtually impossible to answer.

Note also that the answer to such a question must be along the line of a salient set of strategies, policies and/or analogies. E.g. "I would hit the ball like in normal tennis, except I would wait until it bounced higher and aim downwards more." Why? "Gravity no longer brings the ball down as much, so I have to to avoid it going out of bounds. The trajectory will be straighter, so I must hit it at a higher point to avoid the net." It is hard to imagine having such an exchange with a system without language capability, but it's possible that inspecting e.g. an automatic programming system provides us with similarly understandable answers.

3.4 (Re)creation

Creation and recreation of a phenomenon clearly involves some understanding of it. Richard Feynman famously said "What I cannot create, I do not understand". Creation of instances of a phenomenon, often in different situations (e.g. using new materials), typically involves some understanding of what makes the phenomenon what it is—i.e. what are relevant and irrelevant features.

For instance, Figure 1a and Figure 1d both show OR gates, implemented in different ways. An entity that understands NOT (Figure 1c), AND (Figure 1b) and OR gates (Figure 1a), should be able to construct another OR gate out of an AND gate and three NOT gates (Figure 1d). Of course, given these materials, it may also be possible to extract the lower-level electronic components from the given AND and NOT gates and use them to recreate a Figure 1a-style OR gate.

Feynman, like most physicists, wanted to understand the universe. But of course, he did not literally recreate the universe: he created models. To evaluate a system's ability to recreate a phenomenon, we would ideally like to see and interpret the models it has formed. Here we run into a similar problem as with explanation: few if any AI systems can present their models in a way that can be understood and evaluated (by us or themselves).[9]

To evaluate (re)creation abilities of a system, we can give it a design task, where it's given some components / materials and the ability to combine them, and we ask that the phenomenon under test is recreated or that *something* is created that behaves a certain way. Another way is to repeatedly call for increasing levels of compression of the explanation the system gives of a phenomenon — paralleling the progress of science from Earth, Wind, Water & Fire to $E = mc^2$. Clearly, a system that can come up with a compact equation like $E = mc^2$ for some new phenomenon truly understands that phenomenon. This is essentially what Feynman meant.

[9] One interesting case that creates easily interpretable (re)creation models is 3D scene reconstruction [Wojek *et al.*, 2010].

4 Acquiring Understanding

So far we have talked about evaluating the understanding of a certain phenomenon that a purported understander is to be tested on. For general AI a system's ability to acquire understanding of arbitrary phenomena is most likely even more important. In part, we can use the proposed methods to evaluate this ability at separate times, where the system acquires understanding in between. This would give at least two data points indicating how level(s) of understanding change over time; whether and how these measures interact with learning methods (e.g. does the system require tutoring?) and domain (is learning speed dependent on the topic?) will affect how multi-dimensional this measure ultimately is.

We would like to know not only whether an understanding can be acquired by the system, but also how it is acquired: what information and experience is necessary, and how fast this process is. Related questions of importance are how much existing knowledge and understanding is necessary for the system to have before being able to acquire understanding of new phenomena (i.e. what is the maximum delta/time-unit for deepening/broadening understanding?), and what the knowledge transfer function looks like, i.e. does understanding related phenomena help acquiring understanding of new ones, and if so how much?

To evaluate this we need a more general test battery than has been discussed here, and ideally even an entire evaluation framework that allows us to easily construct different (related and unrelated) tasks in order to evaluate the understanding of a multitude of phenomena [Thórisson *et al.*, 2015]. Taking things even further, we would like to know how to teach different phenomena to an AI system, and know that it has received enough information to form an understanding — if it is capable of forming an understanding [Bieger, 2016].

5 Conclusions

We have discussed the need for the evaluation of understanding in intelligent agents, and what tests for this purpose might look like. Four aspects of understanding can be considered separately for this purpose: prediction, goal achievement, explanation and (re)creation. While some of these could be performed successfully by an entity that doesn't understand, it is unlikely that such an entity could succeed on all of these metrics. By viewing phenomena and environments as hierarchical collections of variables with relations between them, we can obtain methods for generating tests for each kind of understanding. Prediction can be tested by querying the state of some variables based on some input information, goal achievement is tested by defining tasks with initial and goal states, explanation is tested by highlighting salient causal events that are most directly responsible for the explained phenomenon, and (re)creation can be tested by formulating a design task or interpreting the system's model of the phenomenon directly.

References

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.

[Besold *et al.*, 2016] Tarek Besold, Stephen Muggleton, Ute Schmid, Alireza Tamaddoni-Nezhad, and Christina Zeller. How does Predicate Invention affect Human Comprehensibility? In *Proceedings of the 26th International Conference on Inductive Logic Programming*. Springer, 2016.

[Bieger, 2016] Jordi Bieger. Artificial Pedagogy: A Proposal. In *Human-Level AI 2016 Doctoral Consortium*, 2016.

[Bostrom, 2014] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[Gentner, 1983] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[Gunning, 2016] David Gunning. Explainable Artificial Intelligence. DARPA-BAA-16-53, 2016.

[Kim, 2015] Been Kim. *Interactive and Interpretable Machine Learning Models for Human Machine Collaboration*. PhD Thesis. Massachusetts Institute of Technology, 2015.

[Legg and Hutter, 2007] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, volume 157, pages 17–24, 2007.

[Levesque *et al.*, 2012] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd Schema Challenge. In *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning*. The AAAI Press, Palo Alto, California, 2012.

[Li *et al.*, 2009] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *IEEE Conference On Computer Vision and Pattern Recognition*, pages 2036–2043. IEEE, 2009.

[Lou *et al.*, 2012] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158. ACM, 2012.

[Nenkova and McKeown, 2011] Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5:103–233, 2011.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference On Computer Vision and Pattern Recognition*, pages 427–436. IEEE, 2015.

[Prager *et al.*, 2000] John Prager, Eric Brown, Anni Coden, and Dragomir Radev. Question-Answering by Predictive Annotation. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[Schmid *et al.*, 2003] Ute Schmid, Helmar Gust, Kai-Uwe Kühnberger, and Jochen Burghardt. An algebraic framework for solving proportional and predictive analogies. In *Proceedings of the First European Cognitive Science Conference*, pages 295–300, 2003.

[Steunebrink *et al.*, 2016] Bas R. Steunebrink, Kristinn R. Thórisson, and Jürgen Schmidhuber. Growing recursive self-improvers. In *Proceedings of the Ninth Conference on Artificial General Intelligence*, pages 129–139. Springer-Verlag, 2016.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.

[Thórisson *et al.*, 2015] Kristinn R. Thórisson, Jordi Bieger, Stephan Schiffel, and Deon Garrett. Towards Flexible Task Environments for Comprehensive Evaluation of Artificial Intelligent Systems & Automatic Learners. In *Proceedings of the Eighth Conference on Artificial General Intelligence*, pages 187–196. Springer-Verlag, 2015.

[Thórisson *et al.*, 2016a] Kristinn R. Thórisson, Jordi Bieger, Thröstur Thorarensen, Jóna S. Sigurðardóttir, and Bas R. Steunebrink. Why Artificial Intelligence Needs a Task Theory — And What it Might Look Like. In *Proceedings of the Ninth Conference on Artificial General Intelligence*, pages 118–128. Springer-Verlag, 2016.

[Thórisson *et al.*, 2016b] Kristinn R. Thórisson, David Kremelberg, Bas R. Steunebrink, and Eric Nivel. About understanding. In *Proceedings of the Ninth Conference on Artificial General Intelligence*, pages 106–117. Springer-Verlag, 2016.

[Wojek *et al.*, 2010] Christian Wojek, Stefan Roth, Konrad Schindler, and Bernt Schiele. Monocular 3D scene modeling and inference: Understanding multi-object traffic scenes. In *Proceedings of the European Conference on Computer Vision*, pages 467–481. Springer, 2010.

[Zelinsky, 2013] Gregory J. Zelinsky. Understanding scene understanding. *Frontiers in Psychology*, 4, 2013.