Formalizing Approximate Objects and Theories: Some Initial Results

Aarati Parmar

Stanford University, Stanford, CA 94305, USA
aarati@cs.stanford.edu,
http://www-formal.stanford.edu/aarati

Abstract. This paper introduces some preliminary formalizations of the approximate entities of [McCarthy, 2000]. Approximate objects, predicates, and theories are considered necessary for human-level AI, and we believe they enable very powerful modes of reasoning (which admittedly are not always sound). Approximation is known as *vagueness* in philosophical circles and is often deplored as a defective aspect of human language which infects the precision of logic. Quite to the contrary, we believe we can tame this monster by formalizing it within logic, and then can "build solid intellectual structures on such swampy conceptual foundations." [McCarthy, 2000].

We first introduce various kinds of approximation, with motivating examples. Then we develop a simple ontology, with minimal philosophical assumptions, in which to cast our formalization. We present our formalization, and show how it captures some ideas of approximation.

1 Introduction

[McCarthy, 2000] introduces approximate objects, predicates, and theories as an extension to AI. Informally, an entity is *approximate* if it can be further refined by finding out more things about it, or by simply defining more. Reasoning with approximate entities ignores unnecessary details, thereby simplifying and accelerating reasoning in general, while remaining somewhat sound. Common sense reasoning will require exactly this property. As in the Aristotle quote of [McCarthy, 2000], "Our discussion will be adequate if it has as much clearness as the subject matter admits of; for precision is not to be sought for alike in all discussions".

We also need to formalize approximation so that we know its boundaries; one needs to know when an approximation fails, and how to move to the next level of precision to reason correctly again. As an example, consider any of the common sense problems displayed in [Morgenstern, 1998], such as cracking an egg. Any theory explaining this process will be inherently approximate, as formalizing every eventuality is tedious and maybe impossible. In general we always feel uncomfortable with any formalization, pointing out its inapplicability with respect to one eventuality or another.[1] The point is, if we can understand how a theory is approximate, then we can accept its failures rather than decry its deficiencies, and simply move on to the next more precise theory when necessary.

Incidentally, many common sense theories of the world use the abnormality predicate Ab and second-order minimization to be robust outside the boundaries of approximation. For example, $\neg Ab(x) \implies (Bird(x) \implies Flies(x))$ formalizes "All birds fly," which, according to the Ab-minimization will infer $Bird(x) \implies Flies(x)$ when consistent to do so. Otherwise it will infer $Ab(x)$, as in the case of a penguin. Other than in action theories and inheritance hierarchies, one never questions how to insert the Abs in a common sense theory, or how to refine them when faced with new information. We hope to provide some preliminary results that will help answer these two questions.

We believe our results may be applicable to other common-sense theories which formalize aspects of the world. Consider any simple grammar describing formation of English sentences ($S \leftarrow NP\ VP, \dots$). It will correctly discriminate a restricted class of sentences, but will have to be increasingly complex as it approaches the full generality of English. Similar phenomena occur for other linguistic theories. However humans, when asked to explain their machinations, give a very simple illustration, adding elaborations only when necessary. This suggests that the formalization of linguistic concepts is not some grand monolithic (and highly complicated) theory but perhaps a series of approximate theories, each expanding and elaborating upon previous theories. This structure would be a lot more elaboration tolerant [McCarthy, 1998] as well.

A contrasting view [Dreyfus and Dreyfus, 1984] asserts that humans do not use rules, but rather discriminate "thousands of special cases." This work describes skill acquisition, the process by which a human masters a domain, as first using a simple set of rules, but ending up with a discriminator of many subtle cases based on experience. The expert does not consciously know these discriminations, and when asked to give rules explaining his behavior, will revert to the basic rules learned as a novice. This explains why an expert system, whose rules are acquired through interviews with experts, are competent but do not perform as well as the experts whose rules it is using. If [Dreyfus and Dreyfus, 1984] describes the true model of human skill acquisition, this would also explain why discriminatory structures such as decision trees, and neural networks have been so popular in AI, as they are models of real-world processes. If this is the case, then at least our theory will formally explain how simple rules get elaborated to a complex structure such as a neural network. If we can interpret the resulting structure as compiled rules, then perhaps our formalism could show how to extract out the declarative versions of these compiled rules. These intriguing avenues are beyond the scope of this preliminary paper.

[1] This is probably the reason why much of AI has focused on toy examples. The examples approximate the world in such a way that minimizes our guilt about how simple the theory is.

1.1 Philosophers' Views on Approximation

We mentioned that what we call approximation, many philosophers in the literature denote as *vagueness*. In general a concept is vague if it has borderline cases, and the boundaries of the indeterminacy are themselves blurred, as are the boundaries of those, and so on [Sorensen, 1997]. Examples given in the literature include baldness, the physical extent of Mount Everest, and the revival of a club (due to Parfit, in [Broome, 1984]). The philosophers debate whether vagueness exists in the world, or is just a linguistic artifact. Any discussion of vagueness involves whether a sharp boundary line truly exists for vague concepts, illustrated by the *sorites paradox*, that set of arguments which states:

1. 1 grain does not make a heap.
2. If n grains do not make a heap, then neither does $n+1$ grains.
3. $\therefore$ 10^6 grains do not make a heap.

Since this argument is false, there must be a point at which n grains are not a heap, but adding 1 more does, which entails that some sharp boundary exists between being a heap and not being a heap.

One of the proposed formalizations of vagueness is *fuzzy logic*, which instead of ascribing true or false values to a sentence, ascribes a value between 0 and 1. Hence if x is only somewhat bald (a borderline case) we might ascribe to the sentence "x is bald" the value 0.5. This value is referred to as the *degree of truth* of the sentence. It seems however, that assigning a vague sentence an absolute value is a bit paradoxical; fuzzy logic avoids this by also allowing *higher types*. That is, instead of assigning a value, one can assign an interval of possible values. Any interval $[a, b]$ can transformed into a fuzzier interval $[[a_1, a_2], [b_1, b_2]]$, where both boundaries of the original interval are fuzzified into intervals themselves. And these can be fuzzified further. If one does not like numbers, one can use any lattice of elements. It turns out that the actual values of numbers are not very important in practice anyway [Goguen, 1968].

Another proposal asserts that vagueness is linguistic, in that some concepts (like baldness) are simply vaguely specified and can have various underlying precise definitions. For example one could define baldness in various ways, as having 0 hairs on one's scalp, having at most 10 hairs, etc. Each one of these possible definitions is a *sharpening* of the concept of baldness, and described with a three-valued logic: a sentence containing a vague concept is true if it holds under all sharpenings of the concept, false if it is false under all sharpenings of the concept, and "indefinite" if it is true for only some of the sharpenings. But then, unfortunately, the law of excluded middle ($\phi \vee \neg\phi$) won't hold for borderline sentences ϕ.[2]

A third view notices that in the previous case, a meta-language is used to give definite truth-conditions for the vaguer target language (although they themselves may be indefinite), and therefore definitely shows at what point truth and falsity dissolve into indeterminacy. So for example in the sorites paradox, at

[2] The argument may be found in [Tye, 2000].

some point whether n grains of sand make a heap will be a truth, and at $n-1$ it will be indeterminate. The fact that there is a sharp line where indeterminacy starts opposes the definition of vagueness. [Tye, 2000] proposes that even the meta-language will need to be vague, as well as the meta-meta-language, etc. This regress is known as higher-order vagueness. (Note that fuzzy logic could model this regress using higher types.)

Finally the *epistemic* view [Williamson, 1994] espouses that definite boundaries for vague concepts exist, but we just cannot know them. Vagueness is an ignorance which we can never overcome.

1.2 Related Work on Abstraction

[Giunchiglia and Walsh, 1992] is one of the first attempts to formalize *abstraction*, defined as the process of mapping the representation of a problem to another one, so that the problem is easier to solve. It also shows how various paradigms, such as ABSTRIPS, [Hobbs, 1985]'s theory of granularity, etc. are instances of various kinds of abstractions. We share the same motivations as [Giunchiglia and Walsh, 1992] in that we both care about the class of mappings that preserve desirable properties while throwing away unnecessary details. However we differ in approach in that we do not concentrate on syntactic mappings between two axiomatic formal systems, which preserve the set of (non-)theorems in some way. While this is important when we discuss *approximate theories*, we care more about the fine structure of such theories, which requires an examination of the nature of objects and predicates in the theory. We want to know what are the intrinsic qualities about theories that make them so amenable to certain forms of abstraction.

1.3 Propositional Approximate Theories

We repeat the treatment of propositional approximate theories given in [McCarthy, 2000]. The ontology includes reality, modeled by a set of propositional variables $r_1, \ldots, r_n$. There are so many observations that can be made, denoted by $o_1, \ldots, o_k$, which are each propositional functions of reality:

$$o_i = \mathcal{O}_i(r_1, \ldots, r_n). \tag{1}$$

The functions model the fact that much of reality is not directly observable. The fact that $n \gg k$ reflects the complexity of reality versus our observations.

$q_1, \ldots, q_l$ are a set of propositions about reality whose values we are interested in learning. They are determined by reality:

$$q_i = \mathcal{Q}_i(r_1, \ldots, r_n). \tag{2}$$

Our theory $\mathcal{AT}$ only approximates $\mathcal{Q}_i(r_1, \ldots, r_n)$ with $\mathcal{Q}'_i(o_1, \ldots, o_k)$, which are only functions of our indirect observations about reality, and not reality itself like the $\mathcal{Q}_i$s.

Lucky$(r_1, \ldots, r_n)$ is a predicate which when true, implies that our approximations to the q_is is correct:

$$\begin{aligned} &Lucky(r_1, \ldots, r_n) \implies \\ &\quad (\forall i \leq l)[\mathcal{Q}_i(r_1, \ldots, r_n) = \mathcal{Q}'_i(\mathcal{O}_1(r_1, \ldots, r_n), \ldots, \mathcal{O}_k(r_1, \ldots, r_n))]. \end{aligned} \tag{3}$$

The main points of this formalization is that:

1. Reality is not always directly observable, partly because of its complexity, and partly because of the sheer enormity of details involved. The idea is that facts about reality may not be *epistemologically adequate*,[3] and the observations o_i are used to express what the subject can in fact sense.
2. The difference in cardinality between n, and k and l also reflects this notion that reality is rich, and our observations are poor. Information is generally lost in the transition from reality to observations. This cardinality of facts is only one way to formalize this idea, and definitely not the most general.
3. Unfortunately, while the q_i are functions of reality, we can only cobble together approximate functions based on our observations. Since our observations are usually lossy compared to reality, it is unlikely that the $\mathcal{Q}_i$ and $\mathcal{Q}'_i$ will coincide. (Only if we are *Lucky*!)

1.4 The Ontology of Approximate Things

Here we further interpret the approximate objects, predicates, and theories introduced in [McCarthy, 2000]. An *approximate object* is an object o in a logical theory T, which either only partially captures the properties of some real object in *reality* or is itself inherently partial. The approximateness is inherent to the object, and not a matter of incompleteness of the theory (although to maintain consistency a theory with approximate objects may have to abandon completeness). Rather it is more a fact that the theory or its language may not be able to properly capture an object's properties. We can attempt a definition of approximate objects by first expounding three kinds of approximations of objects that occur in common sense reasoning about the world:

1. (TYPE I) The first kind of approximate object is that which represents an epistemologically richer[4] object. For example, most Blocksworld theories idealize a block so extremely as to ignore its physical characteristics, as well as other relevant properties, such as its mass. Such an approximation generally ignores irrelevant properties as well, such as the color of the block, or where it was manufactured. It is clear that this sort of abstraction is useful

[3] "[McCarthy and Hayes, 1969] defines an epistemologically adequate representation of information as one that can express the information actually available to a subject under given circumstances." from [McCarthy, 1979]

[4] [Howe, 1994] describes *rich* objects as those for which can be asserted properties, which we cannot be completely described. *Poor* objects are exactly the opposite.

in cases, to strain out many [irrelevant] facts, to simplify the ontology to the task at hand. Approximations are most useful when an epistemologically rich object is approximated by a poor one, to eliminate an entire order of complexity.

2. (TYPE II) Another kind of approximation is the mental conception of objects for which there is no basis in reality. One example includes the concept of being middle-aged. This is an abstraction constructed by humans, which has no corresponding concept in reality: there are facts about being "middle-aged" for which there is no inherent truth, and can only be decided by *defining more*.
   Another example is the concept of being red. In reality, there is no such concept; there are only wavelengths, generally between 600 and 700 nanometers, that give us humans the sensation of seeing red. Baldness, and what a heap is, are other such approximations, and the paradoxes that arise from studying them too closely arise from the fact that there is no real corresponding concept in reality that would decide the matter.
   This kind of approximation can be thought of as a human-created characterization of some phenomenon, which either simplifies reasoning or organizes it nicely. These kinds of approximation can be used to characterize a wide range of categories, without wasting too many words on a complete specification. Because they are not completely specified, the definitions will be incoherent with respect to reality, often leading to paradoxes like that of the heap. Since the definitions are defined by humans they are subject to cultural as well as personal biases.
3. (TYPE III) The final type of approximate objects are those that arise in counterfactual sentences [Costello and McCarthy, 1999]. Like Type II objects, they have no direct analogy in reality, by definition. They are only defined by whatever properties that are ascribed to them in the counterfactual. There is no truth of the matter for any other properties not mentioned in (or derivable from) the counterfactual.
   For example in "If another car had come over the hill when you passed that car, there would have been a head-on collision," the other car could have been a Buick, a Mercedes, etc. There is no truth to the matter about what make of car it was.

An *approximate predicate* is one whose extent is vague or ill-defined. There are borderline cases whose membership is questionable, and therefore it is difficult to come up with necessary and sufficient conditions. Some illustrative examples include "the wants of the U.S." [McCarthy, 2000] and religion [Alston, 1967]. We define *concepts* as unary predicates, which can be approximate, such as natural kinds (is an orange lemon still a lemon?). Fuzzy logic has had success in this area, as it allows one to talk about the *degree of membership* of an object in a set, which can represent borderline cases.

We hope to address the relation between our formalism and fuzzy logic in later work. Our intuition is that a logically defined concept ϕ will be vague because it is fundamentally ill-defined or incoherent, such as baldness or heaps.

Fuzzy logic can handle the questionable cases by ascribing partial degrees of truth. In first order logic, the only recourse is to either have an incomplete theory, in the formal sense that for some object c, $T \not\models bald(c) \wedge T \not\models \neg bald(c)$, or to restrict the theory to some limited domain of discourse. (such as the set of people who are either definitely bald or not.) A much better approach would explicate how and when a concept is ill-defined, with respect to a more accurate theory. This would then be superior to the fuzzy logic approach, since it would explain *why* borderline cases are borderline, rather than sweeping the problem under the rug by ascribing a partial degree of truth.

An *approximate theory* is a set of sentences which is an abbreviated description (in some sense) of some phenomenon. One example is when unnecessary details about the world are ignored: reasoning about what will be served for lunch on a flight is not required to plan a trip to Hawaii. Another is a kind of idealization such as Blocksworld. We give a framework in which to talk about theory approximation in §3.

The rest of this paper follows this outline: after some mathematical notation is introduced (§2), we delve into a proposed ontology for first order theories (§3), and then a formal definition of when one theory approximates another (§4), with some examples. §5 formalizes epistemologically rich and poor objects, which enables us to address the different types of approximate objects (§6). Finally in §7 we consider when an approximate theory is *coherent* (correctly predicts reality) and conclude in §8.

2 Preliminaries

We describe our notation and simplifications here. $\mathfrak{A}$, $\mathfrak{B}$, $\mathfrak{R}$, $\mathfrak{M}$, and $\mathfrak{M}'$ are first-order structures with non-empty universes. The universe of $\mathfrak{A}$ is denoted $|\mathfrak{A}|$. $\mathbf{M}$ denotes a class of first-order structures. T, T_A, T_B are all consistent first-order theories. Any symbol of the form $\mathcal{L}_{\mathcal{X}}$ is a signature, or language.

We often interchange, in proofs and examples, a class of models $\mathbf{M}$ with the theory T describing them. This should not be a source of confusion, as we know there are well-defined functions $Th_{\mathcal{L}}(\mathbf{M})$ which given a class of models in language $\mathcal{L}$ returns the set of formulas true in all of them, and $Mod(T)$ which returns the class of models of T.

Finally, if X is a set, then X^n is the set of n-tuples of X. $\mathfrak{A} \cong \mathfrak{B}$ means that there is an isomorphism between the structures $\mathfrak{A}$ and $\mathfrak{B}$. If two models are isomorphic then they entail the same set of sentences.

3 Our Ontology

In this section we would like to construct a formalism which reflects the intuitions in §1.3, but is more general, and allows first-order statements rather than propositions. To this end, we introduce the following ontology:

Let $\mathfrak{R}$ be a first-order model of (one's idea of) reality, using some language $\mathcal{L}_{\mathcal{R}}$. We let $\mathbf{R}$ be the collection or class of such models. On the other hand, our observations are constructed in the language $\mathcal{L}_\mathcal{O}$. The language $\mathcal{L}_\mathcal{O}$ is determined by the observations that can be made by our sensory organs. The standard set of philosophical inquiries that we could make about the existence of the models of reality $\mathbf{R}$ is moot, since each element $\mathfrak{R} \in \mathbf{R}$ is supposed to be the observer's *perception* or idealization of what reality is, which may be isomorphic to reality, but not necessarily so. So for example, to a particle physicist, $\mathfrak{R}$ would include interactions between quarks. The $\mathfrak{R}$ of an ancient Greek would explain weather patterns in terms of Zeus' temper and throwing lightning bolts.

While $\mathcal{L}_\mathcal{O}$ must be *epistemologically adequate*, $\mathcal{L}_\mathcal{R}$ should be *metaphysically adequate*.[5] We use a class of models instead of a sole model to represent different perceptions of reality (wave versus particle interpretation of light), or incompleteness in one's perception of reality. If there is no such incompleteness, then $\mathbf{R}$ can just be $\{\mathfrak{R}\}$, the singleton model.

The relation between $\mathcal{L}_\mathcal{O}$ and $\mathcal{L}_\mathcal{R}$ not only explains how objects are mapped from one's sensations to reality (in other words, how they are grounded), but impart structure to any theory of reality based on our observations. $\mathcal{L}_\mathcal{R}$ describes very rich phenomena, which explains why $\mathcal{L}_\mathcal{O}$, which is meant to be a relatively simple language, does not coincide with it. Note that instead of providing an axiomatization of $\mathbf{R}$, we provide the models themselves. This is because there may not be a finite first-order axiomatization of $\mathbf{R}$. We also prefer classes of models to a theory, because assaying truth in reality appears to be more of a determination of whether it holds in the world (whether $\mathfrak{R} \models \phi$), rather than some process of inference based on statements (whether $\mathfrak{R} \vdash \phi$).

4 Theories Approximating Theories

The method of *syntactic interpretation* used by [Tarski et al., 1953] used to prove [un]decidability of theories can be used to express whether one theory approximates another. We copy the presentation in [Baudisch et al., 1985]: given two classes of models $\mathbf{A}$ and $\mathbf{B}$, respectively, of languages $\mathcal{L}_\mathcal{A}$ and $\mathcal{L}_\mathcal{B}$, we define an interpretation $I : \mathcal{L}_\mathcal{A} \to \mathcal{L}_\mathcal{B}$ which sends every predicate symbol $P(\overline{x})$ of $\mathcal{L}_\mathcal{A}$ to the formula $\phi_P(\overline{x})$ in $\mathcal{L}_\mathcal{B}$, and the formula $x = x$ to $\phi_=(x)$, which is a formula of $\mathcal{L}_\mathcal{B}$. For now we assume that we only have relational symbols, and no functions or constants in $\mathcal{L}_\mathcal{A}$, as they can be represented in terms of relations and some extra axioms. We explain the transformation for constant and function symbols in §4.1.

This interpretation is extended to all formulas in $\mathcal{L}_\mathcal{A}$ using the following inductive rules, where α and β are formulas of $\mathcal{L}_\mathcal{A}$, x and y variables, and $\overline{x}$ a tuple of variables:

1. $(x = y)^I = x = y$
2. $(P(\overline{x}))^I = \phi_P(\overline{x})$

[5] [McCarthy, 1979] defines metaphysically adequate representations as those “that can represent complete facts ignoring the subject's ability to acquire the facts in given circumstances.”

3. $(\neg\alpha)^I = \neg(\alpha)^I$
4. $(\alpha \vee \beta)^I = \alpha^I \vee \beta^I$
5. $(\exists x \alpha)^I = (\exists x)(\phi_=(x) \wedge \alpha^I)$

From any model $\mathfrak{B} \in \mathbf{B}$ in the target language $\mathcal{L}_\mathcal{B}$, we can define a structure $\mathfrak{B}^I$ in $\mathcal{L}_\mathcal{A}$. This consists of two simple steps:

1. We simply set the universe of $\mathfrak{B}^I$ to be the set of elements in $\mathfrak{B}$ obeying $\phi_=(x)$:

$$|\mathfrak{B}^I| = \{x \in |\mathfrak{B}| \mid \mathfrak{B} \models \phi_=(x)\}. \tag{4}$$

2. Then we map the interpretation of each n-ary predicate P of $\mathcal{L}_\mathcal{A}$:

$$P^{\mathfrak{B}^I} = \{\overline{x} \in |\mathfrak{B}^I|^n \mid \mathfrak{B} \models \phi_P(\overline{x})\}. \tag{5}$$

   For each predicate $P \in \mathcal{L}_\mathcal{A}$, $\mathfrak{B}^I$ "back-translates" a possible definition for it by looking at ϕ_P in $\mathfrak{B}$. P's domain of application is controlled by the predicate $\phi_=$, which picks out the part of $\mathfrak{B}$ corresponding to objects in T_A.

One can finally define a notion of *interpretability*:

Definition 1 (Interpretability). *Let $T_A = Th_{\mathcal{L}_\mathcal{A}}(\mathbf{A})$ and $T_B = Th_{\mathcal{L}_\mathcal{B}}(\mathbf{B})$. Then the theory T_A is* interpretable *in T_B, or T_B* interprets *T_A if there is an interpretation function $I : \mathcal{L}_\mathcal{A} \to \mathcal{L}_\mathcal{B}$ such that:*

1. *for every structure $\mathfrak{B} \in \mathbf{B}$ there is a $\mathfrak{A} \in \mathbf{A}$ such that $\mathfrak{B}^I \cong \mathfrak{A}$ and,*
2. *for every structure $\mathfrak{A} \in \mathbf{A}$ there is a $\mathfrak{B} \in \mathbf{B}$ such that $\mathfrak{B}^I \cong \mathfrak{A}$.*

The two requirements for interpretability can be explained as (1): every model of T_B can be "back-translated" by I to be a model of T_A, and (2): every model of T_A can be expanded to be a model of T_B. A small theorem shows us how strong the concept of interpretability is:

Theorem 1 (Interpretability of formulas). *Assume T_B interprets T_A. Then for any $\phi \in \mathcal{L}_\mathcal{A}$,*

$$T_A \models \phi \iff T_B \models \phi^I \tag{6}$$

Proof. *We include the lemma:*

Lemma 1 ([Rabin, 1965]). *Given any $I : \mathcal{L}_\mathcal{A} \to \mathcal{L}_\mathcal{B}$, for any such $\phi \in \mathcal{L}_\mathcal{A}$, and structure $\mathfrak{B}$ of $_\mathcal{B}$,*

$$\mathfrak{B} \models \phi^I \iff \mathfrak{B}^I \models \phi. \tag{7}$$

This lemma is proved by induction on the complexity of formulas.

$\rightarrow$: Assume $T_A \models \phi$. Let $\mathfrak{B} \models T_B$, and try to show $\mathfrak{B} \models \phi^I$. By the definition of interpretability, there is an $\mathfrak{A}' \models T_A$ such that $\mathfrak{B}^I \cong \mathfrak{A}'$. Hence $\mathfrak{B}^I \models T_A$, which means that $\mathfrak{B}^I \models \phi$. By the lemma $\mathfrak{B} \models \phi^I$.

$\leftarrow$: This time assume $T_B \models \phi^I$, let $\mathfrak{A} \models T_A$, and show $\mathfrak{A} \models \phi$. By interpretability there is a $\mathfrak{B}' \models T_B$ such that $\mathfrak{A} \cong {\mathfrak{B}'}^I$. $\mathfrak{B}' \models \phi^I$ and by the lemma this means ${\mathfrak{B}'}^I \models \phi$. And then from the isomorphism we can deduce $\mathfrak{A} \models \phi$.

Hence if T_B interprets T_A, for any formula $\phi \in \mathcal{L_A}$, we can use T_B to find out whether $T_A \models \phi$ simply by checking if the interpreted formula ϕ^I holds in T_B. Intuitively, the function I interprets a simple theory T_A in a more complicated on T_B; T_B contains the concepts necessary to express the concepts formalized in T_A.

We can now define the simplest form of theory approximation. We can say that a theory T_A in language $\mathcal{L_A}$ *approximates* another theory T_B in $\mathcal{L_B}$, written $T_A <_{approx} T_B$, if T_A is interpretable in T_B but not the other way around. The idea is that T_B is powerful enough to model T_A, but T_A cannot do the same for T_B, so it must be inherently more approximate (lost information).

This method of syntactic interpretations gives us a framework in which to relate different theories with different languages through the interpretation function I. In order to make further distinctions in different kinds of approximation we believe it will be necessary to study the fine structure of the function I. I may be related to the *simplifying assumptions* mentioned in [Nayak and Levy, 1995].

Some facts to notice before we continue with some examples: Assume T_B interprets T_A. If T_B were complete, then it has only one model $\mathfrak{B}$, and (2) means that all the models of T_A are isomorphic to each other. Therefore T_A is categorical, and complete as well. If T_B is complete, then any approximation to it, according to this definition will have to be as well. On the other hand, if T_A were complete, then every model of T_B will have to agree when back-translated to $\mathcal{L_A}$.

4.1 The Treatment of Functions and Constants under Approximation

This section illustrates how functions and constants in $\mathcal{L_A}$ are transformed under the function I, and where they appear in the models of T_A versus models of T_B. Assume f is a unary function symbol of $\mathcal{L_A}$, used in T_A. We can construct a predicate $P_f(x, y) \iff_{def} f(x) = y$ which represents the *graph* of f, and have a new language $\mathcal{L}'_\mathcal{A} = \mathcal{L_A} - \{f\} + \{P_f\}$. Also we alter our theory $T'_A \equiv T_A|_{f(x)=y \leftarrow P_f(x,y)} \wedge (\forall x)(\exists y)(\forall z)[P_f(x, z) \iff y = z]$. Then the translation would be:

$$((\forall x)(\exists y)(\forall z)[P_f(x, z) \iff y = z])^I = \\ (\forall x \in \phi_=)(\exists y \in \phi_=)(\forall z \in \phi_=)[\phi_{P_f}(x, z) \iff y = z]. \tag{8}$$

By Theorem 1, we know that $T_B \models (8)$, so that ϕ_{P_f} defines within T_B a function (call it g), which bears a relation to f through the transformation from P_f and ϕ_{P_f}. Also, g is only defined on the extension of $\phi_=$.

Since a constant is a 0-ary function, we can recycle the exposition above to show that our graph of $c(x) = c$ is $P_c(y)$, and (8) will assert that P_c is the predicate uniquely true of c. Then under I we will get another predicate ϕ_{P_c} which will be uniquely true of some other element $d \in \phi_=$. Intuitively then we can consider I to also map objects and functions, where we denote $d = c^I$, and $g = f^I$, which are related through the predicate transformations.

4.2 Theory Approximation Example: Generalizing Conditionals

Consider the theory $T_B \equiv \Psi \wedge (\forall x)[P(x) \wedge \alpha(x) \implies \beta(x)]$, where P is a complicated predicate we would like to ignore, and Ψ some set of sentences which does not mention P. Can we approximate it by the much simpler theory $T_A = \Psi \wedge (\forall x)[\alpha(x) \implies \beta(x)]$?

An obvious interpretation is if $[P(x)]^I = \phi_=(x)$ – that is, T_A can be interpreted in any model of T_B provided we restrict the domain of discourse to objects obeying $P(x)$. There could be other interpretations that depend on the structure of α, β and Ψ. The details are omitted here.

4.3 Theory Approximation Example: Blocksworld

Let T_{BW} be the standard situation calculus theory of Blocksworld, using the language $\mathcal{L}_{\mathcal{BW}} = (On(x,y,s), move(x,y,z), Table, A, B, C, Result)$ where $On(x,y,s)$ is the predicate stating block x is on block y, $move(x,y,z)$ is the action that moves x from y to the top of z, $Table$ denotes a table of infinite capacity, and A, B, and C are names for unique blocks. *Result* is the standard successor function used in situation calculus. We assume T_{BW} has successor state axioms to completely specify the effects of every action; if an action is not possible, we assume the world stays as it is. As we mentioned before, this theory is so simple it does not even model the physical aspect of the blocks. Another important approximation is that the table has infinite capacity. Finally, blocks are either on, or off another block – there is no concept of them being partially on a block, or being on two blocks at the same time.

Now consider a more realistic theory T'_{BW} of Blocksworld that models the same blocks as in T_{BW}, but as roughly parallelpieds, each with a center of gravity. If the center of gravity of any tower of blocks is not supported from below, this will generate torque about this axis and cause the blocks to fall to the table. The language $\mathcal{L}'_{\mathcal{BW}}$ for such a descriptive theory includes symbols such as: $On(x,y,\delta,s)$, where δ is the deviation of the center of gravity of x from y, and $move(x,y,z,\delta)$, which is the action of moving x from y to z with δ deviation of x's center of gravity from z's. Other than the constants $Table$, A, B, and C, are functions $cg(x)$ which gives the center of the gravity of the tower of blocks above and including x, and normal arithmetic functions. Of course we need some function $surface(y,s)$, which returns some structure delineating the surface of y (so that we can check if a tower of blocks is not supported and will fall).

We argue that $T_{BW} <_{approx} T'_{BW}$. To show this, we must show that T'_{BW} interprets T_{BW}, but not the other way around. Consider the interpretation I

which interprets $On(x, y, s)$ as the formula $(\exists\delta)On(x, y, \delta, s)$, and $move(x, y, z) = a$ as $(\exists\delta)[a = move(x, y, z, \delta) \wedge \delta + cg(x) \in surface(z, s)]$. We also set $\phi_=(x)$ to be true only for the blocks and the table, and for actions which place blocks precisely over the center of gravity of other blocks. Each block A, B, C is matched to the corresponding block in T'_{BW}, and the $Table$ to the $Table'$. The idea behind this interpretation is that models of T_{BW} should correspond to "rounded-off" versions of models of T'_{BW}.

Take any model $\mathfrak{M}$ of T_{BW}. We must show there is a corresponding $\mathfrak{M}'$ of T'_{BW} such that ${\mathfrak{M}'}^I$ and $\mathfrak{M}$ are isomorphic. We construct a model $\mathfrak{M}'$ of T'_{BW} by taking $|\mathfrak{M}'| \supseteq |\mathfrak{M}|$. Then for each $\langle x, y, s\rangle \in On^{\mathfrak{M}}$, we put $\langle x, y, 0, s\rangle \in On^{\mathfrak{M}'}$. If $a = move^{\mathfrak{M}}(x, y, z)$, then we set $s = move^{\mathfrak{M}'}(x, y, z, 0)$. Clearly by definition $\mathfrak{M} = {\mathfrak{M}'}^I$, since $|\phi_=| = |\mathfrak{M}|$.

For the second condition, it is enough to show that for any model $\mathfrak{M}'$ of T'_{BW}, ${\mathfrak{M}'}^I \models T_{BW}$. $\phi_=$ in this case will make sure $|{\mathfrak{M}'}^I|$ consists only of blocks, the table, and precise moves. Then every "inexact" placement of blocks $On(x, y, \delta, s)$ is rounded down to an exact one in ${\mathfrak{M}'}^I$, and every "inexact" movement $move(x, y, z, \delta)$ becomes the exact $move(x, y, z)$.

On the other hand, T_{BW} does not interpret T'_{BW}. T'_{BW} is simply more expressive. Formally, we can show this by considering the contrapositive of Theorem 1, by taking an arbitrary $I : \mathcal{L}'_{\mathcal{BW}} \to \mathcal{L}_{\mathcal{BW}}$, and finding a ϕ which does not translate over. The trick is to have ϕ formalize one of the differences between T'_{BW} and T_{BW}. For example, ϕ could be the statement that there exists a move where the block falls to the table instead of reaching its destination (because it is placed haphazardly on another block). $T'_{BW} \models \phi$. But there is no way to translate this to an expression in $\mathcal{L}_{\mathcal{BW}}$ which talks about the mysterious failure of an action, without contradicting T_{BW}.

4.4 Related Work

[Nayak and Levy, 1995] uses the same mathematical framework of syntactic interpretation to characterize *model increasing (MI) abstractions*. MI abstractions have the advantage over the syntactic ones in [Giunchiglia and Walsh, 1992] in that they capture more of the "underlying justification that leads to the abstraction," [Nayak and Levy, 1995]. Among other insights, the work shows how the ABSTRIPS abstraction is an MI one. [Levy, 1994] formalizes *irrelevance* of clauses with respect to queries on knowledge bases, as well as independence of predicate arguments. [Nayak, 1994] combines abstractions within the theory of contexts.

5 Rich and Poor Objects

A *rich* object is one that cannot be completely described, while a *poor* one can. At first, one possible criterion for "description" may be identifying properties, the set of properties which uniquely specify the object. This is opposed to *all* of the properties which are true of the object. The number 0, even though there are infinitely many things to be said of it $(0 < 1, 0 < 2, 0 < 3, 0 < 4, \ldots)$, can be concisely identified as the unique number which has no predecessor, and therefore is poor. On the other hand, the author may be uniquely identified, by the poor description "the person writing this paper," but the author is a rich object, not a poor one.

Whether an object is rich or poor also depends on how it is formalized, and what language is used to represent it. Situations, defined as a "snapshot of the world," are always treated as rich objects, while states, a finite collection of variables' values, are poor. A situation is rich because we can always find another property that specifies the situation further, while once we decide upon n variable values, a state is completely specified and very unmysterious. But then if n were very large, and only a small part of the state's variable values were ever contemplated in a theory, it might as well be a rich object.

A possible definition of a rich object is one for which we can always extend a theory of it to one which ascribes more untrivial properties. If o_R is a rich object in theory T, then we can imagine a more expressive theory T' where $T <_{approx} T'$ in which o_R^I appears. Any concept true of o_R in T will have its interpretation true of o_R^I in T' by Theorem 1, and if the interpretation I is injective, the distinct properties true of o_R can only increase. But to guarantee that we find out more interesting facts about o (and not some other object in T') we will have to require something like that the $tp_{T'}(o_R^I)$, the *type* of o_R^I, defined as the set of all 1-place formulas of T' that are true of o_R^I, is not interpretable in T.

A *rich* object o_R in theory T_0 then, would be one for we can continually de-approximate theories about o_R: there is an infinite sequence of theories such that $T_0 <_{approx} T_1 <_{approx} T_2 <_{approx} T_3 \ldots$. Therefore an object o_P is *poor* iff every such sequence of theories bottoms out; we run out of things to say. One object o is *richer* than another o' if the longest sequence of o' theories is less than the longest one of o's, starting from the simplest theory $o = o$ or $o' = o'$. Back to the state example, a state with 10^7 variables is richer than one with 3, but both are poor, and can't compare to a situation.

We could define reality $\mathbf{R}$ as the class of models which bound every such de-approximating sequence of theories: for every extension of a theory in $\mathcal{L}_{\mathcal{O}}$, the supremum of the sequence is a theory T_{sup} whose models are contained in $\mathbf{R}$.

6 Back to the Objects

In §3 we presented a mathematical definition for what it meant for one theory to approximate another. Now we return to our task of formalizing each of the three types of approximate objects within the theory. The formalization of a theory approximation is needed for object approximation because as we noted earlier, whether an object is approximate depends on its context, which refers to how it is described and what language is used to describe it.

In the description of these types of objects in §1.4, we compared each object in some given theory (assumed to be in language $\mathcal{L}_{\mathcal{O}}$) to its counterpart in reality in $\mathcal{L}_{\mathcal{R}}$. Below instead of talking about theories in $\mathcal{L}_{\mathcal{O}}$ and $\mathcal{L}_{\mathcal{R}}$, we generalize the discussion to theories T_A and T_B, where $T_A <_{approx} T_B$. Thus the definitions below will apply to any pair of theories.

6.1 Approximate Objects of Type I (Poorer Version of a Richer Object)

Let c be an object constant, in theory T_A, of type I. This means that there exists a more precise theory T_B with corresponding object c^I, for which c is but a poor approximation. This is represented by three facts:

1. $T_A <_{approx} T_B$.
2. c has some corresponding object, c^I in T_B.
3. The description of c in T_A is poorer than that of c^I in T_B.

To implement type I objects all we require is a de-approximation T_B to the current theory T_A. The existence of c^I will be guaranteed by the interpretation function I, and since generally I is injective, will map different properties in T_A to different ones in T_B, so that only more things can be said of c^I in T_B. Naturally, we must add the constraints given in §5 about $tp_{T_B}(c^I)$ not being expressible in T_A to make sure its description gets richer.

If $T_B = Th_{\mathcal{L}_{\mathcal{R}}}(\mathbf{R})$, almost every object o we imagine would be approximate, since every theory T of o can be extended to $Th_{\mathcal{L}_{\mathcal{R}}}(\mathbf{R})$ (it being a $<_{approx}$ upper bound on all theories), and we can always imagine o's richer analog in reality.

6.2 Approximate Objects of Type II (Objects with no Basis in Reality)

If an object (or concept) c has no basis in reality, then of course an interpretation I cannot be built, since c maps to some concept c^I in $\mathfrak{R}$, which cannot exist! The color red, "what the U.S. wants," corners, baldness, and heaps are all examples. Although these concepts have no immediate corresponding object in $\mathbf{R}$, they do seem to correspond to some composition of objects in reality. For example, "red" is the sensation of viewing light of wavelengths from 600 to 700 nanometers. Similarly, "what the U.S. wants" is some complicated array of what the President, U.S. diplomats, populace, etc. desire. A corner is some spatial extent about a physical corner, while baldness is some collection of states of a person with very little hair. The objects are approximate not only because they correspond to a composition of objects, but the nature of composition is itself vague.

For many of these examples, the approximation is an association of one object in our base theory T_A with a *collection* of objects in a more precise theory $T_A <_{approx} T_B$. For example the concept of a *heap* in T_A is associated with some set of actual heaps in T_B. We need to have an altered interpretation function that maps objects in T_A to sets in T_B. The epistemology of this is not worked out yet and we hope to say much more later.

6.3 Approximate Objects of Type III (Counterfactual Objects)

[McCarthy, 2000] introduces a function $Whatif?(p, x)$, where p is a proposition, and x some constant symbol. It can be used to consider counterfactual concepts such as "What would have happened if another car had come over the hill when you passed that Mercedes." We can adapt this interesting function and consider $Whatif?(p, T)$, where T is a first-order theory about the world and p still a proposition, both in the same language $\mathcal{L}$. $Whatif?(p, T)$ would be the resulting theory if p were to hold.

Clearly, $Whatif?(p, T) \models p$. Also, if $T \models p$, then $Whatif?(p, T) = T$. If this isn't the case, $Whatif?(p, T)$ can be imagined as some minimal re-arrangement of T that would be consistent with p. $Whatif?$ could be implemented using the mechanisms in [Costello and McCarthy, 1999]. For non-trivial p and T, $Whatif?(p, T)$ could be infinitely refinable, which means that the function itself would be rich, and any finite theory it returned would only be some approximation! To try to keep this from happening we assume that $Whatif?(p, T)$ returns a theory also in language $\mathcal{L}$. Hence the more expressive $\mathcal{L}$ is, the more interesting $Whatif?(p, T)$ will be.

$Whatif?(p, T)$, being a counterfactual theory, will reference objects of type III, which won't exist in T, as in "the car that came over the hill." The properties of these type III objects should be limited to those ascribed to them by p, along with whatever ramifications that may follow from $Whatif?(p, T)$.

7 When an Approximate Theory Works in Reality

Consider a robot stacking blocks. Suppose all the blocks are very close to being idealized cubes, and the servo-mechanism in the robot arm is programmed so carefully that the robot always stacks blocks precisely without error. Then to the robot, the typical theory of Blocksworld will be a true description of reality, even though we smarter humans know that it is a highly idealized description of the world that won't work in general.

But to the robot, this theory is correct. We would like to define under what circumstances a simple approximate theory will work correctly in that complex reality, because then we will know when we can get away with such approximations. This may also give us hints as how to parameterize a theory with *Ab*s so that it will be more robust outside the boundaries of approximation, and also more elaboration tolerant.

We define this idea of when an approximate theory correctly predicts reality as *coherence*. A theory T is *coherent* with respect to a more realistic theory T' and restrictions (same as simplifying assumptions) Φ if: I is an interpretation from $\mathcal{L}$ to $\mathcal{L}'$ and $T' \wedge \Phi \models T^I$, that is, every model of $T' \wedge \Phi$ will also model T interpreted in T'. If $T <_{approx} T'$ then coherence follows. Coherence is *Lucky*-ness for first order theories.

There are already examples of coherent theories in this paper. In §4.2, if we restrict our domain of discourse to be $\{x \mid P(x)\}$ we can approximate $\Psi \wedge$

$(\forall x)[P(x) \wedge \alpha(x) \implies \beta(x)]$, with $\Psi \wedge (\forall x)[\alpha(x) \implies \beta(x)]$. This is obvious, but we can imagine more complicated theories for which we would need the formalism to deduce when approximations would apply. If $P(x) = \neg Ab(x)$ we could apply the theory in reverse to find out how to add *Ab*s to a theory to make it more elaboration tolerant.

8 Conclusions

This paper is very preliminary and many of the definitions of approximation are yet only approximations themselves. Much more work needs to be done in order to make our theories precise. We need to further explore the structure of the interpretation functions I and understand how they map concepts in one theory to those in another, especially with regard to approximate objects of type II. This will require an understanding of how to ground symbols. Perhaps we can develop this understanding by considering the approximation of $T_A <_{approx} T_B$ only within the context of how they relate to reality $\mathbf{R}$.

Some other formal ideas from mathematics may be relevant to the properties of epistemologically rich and poor objects. For example forcing and generic sets [Feferman, 1965] formalize the notion that a concept can be described in some finite set of sentences, which could be a better criterion for whether an object is poor. The theory sequences we use to determine richness/poorness might be related to the sequence of finite sets of conditions $Q_0, Q_1, \ldots$.

At least this theory explains why there are so many different AI theories of action and change. Each formalism is just a different approximation to reality. $<_{approx}$ is not a total order, so many different, incomparable approximations are possible. Also, an initial review of [Fine, 1985] suggests that *arbitrary objects* are a kind of approximate object that would be fruitful to study.

9 Acknowledgments

We are thankful for guidance from John McCarthy, as well as Rada Chirkova and Sheila McIlraith for pointers to additional work in abstraction. We are also grateful to the anonymous referees for their helpful comments and suggestions.

References

[Alston, 1967] Alston, W. P. (1967). Vagueness. In Edwards, P., editor, *Encyclopedia of Philosophy*, volume 8, pages 218–221. MacMillan.

[Baudisch et al., 1985] Baudisch, A., Seese, D., Tuschik, P., and Weese, M. (1985). *Model-Theoretic Logics*, chapter Decidability and Quantifier Elimination, pages 235–268. Springer-Verlag.

[Broome, 1984] Broome, J. (1984). Indefiniteness in identity. *Analysis*, 44:6–12.

[Costello and McCarthy, 1999] Costello, T. and McCarthy, J. (1999). Useful Counterfactuals[6]. *Electronic Transactions on Artificial Intelligence*.

[6] http://www-formal.stanford.edu/jmc/counterfactuals.html

[Dreyfus and Dreyfus, 1984] Dreyfus, H. and Dreyfus, S. (1984). From Socrates to Expert Systems: The Limits of Calculative Rationality[7].

[Feferman, 1965] Feferman, S. (1965). Some applications of the notions of forcing and generic sets. *Fundamenta Mathematicae*, 56:325–345.

[Fine, 1985] Fine, K. (1985). *Reasoning with Arbitrary Objects*. Aristotelian Society Series. Oxford.

[Giunchiglia and Walsh, 1992] Giunchiglia, F. and Walsh, T. (1992). A theory of abstraction. *Artificial Intelligence*, 57(2-3):323–389.

[Goguen, 1968] Goguen, J. A. (1968). The logic of inexact concepts. *Synthese*, 19:325–373.

[Hobbs, 1985] Hobbs, J. R. (1985). Granularity. In *International Joint Conference on Artificial Intelligence (IJCAI'85)*, pages 432–435.

[Howe, 1994] Howe (1994). Rich Object[8]. In Howe, D., editor, *Free Online Dictionary of Computing*.

[Levy, 1994] Levy, A. Y. (1994). Creating abstractions using relevance reasoning. In *AAAI, Vol. 1*, pages 588–594.

[McCarthy, 1979] McCarthy, J. (1979). Ascribing mental qualities to machines[9]. In Ringle, M., editor, *Philosophical Perspectives in Artificial Intelligence*. Harvester Press. Reprinted in [McCarthy, 1990].

[McCarthy, 1990] McCarthy, J. (1990). *Formalizing Common Sense: Papers by John McCarthy*. Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648.

[McCarthy, 1998] McCarthy, J. (1998). Elaboration Tolerance[10]. In *In Proceedings of the Fourth Symposium on Logical Formalizations of Common Sense Reasoning*.

[McCarthy, 2000] McCarthy, J. (2000). Approximate objects and approximate theories. In Cohn, A. G., Giunchiglia, F., and Selman, B., editors, *KR2000: Principles of Knowledge Representation and Reasoning,Proceedings of the Seventh International conference*, pages 519–26. Morgan-Kaufman.

[McCarthy and Hayes, 1969] McCarthy, J. and Hayes, P. J. (1969). Some Philosophical Problems from the Standpoint of Artificial Intelligence[11]. In Meltzer, B. and Michie, D., editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press.

[Morgenstern, 1998] Morgenstern, L. (1998). Common Sense Problem Page[12]. Webpage.

[Nayak, 1994] Nayak, P. P. (1994). Representing multiple theories. In Hayes-Roth, B. and Korf, R., editors, *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pages 1154–1160, Menlo Park, CA. AAAI Press.

[Nayak and Levy, 1995] Nayak, P. P. and Levy, A. (1995). A semantic theory of abstractions. In Mellish, C., editor, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 196–203, San Francisco. Morgan Kaufmann.

[Rabin, 1965] Rabin, M. O. (1965). A simple method for undecidability proofs and some applications. In Bar-Hillel, Y., editor, *Logic, Methodology and Philosophy of Science*, pages 58–68. North Holland Publishing Company.

[7] http://socrates.berkeley.edu/~hdreyfus/html/paper_socrates.html
[8] http://burks.brighton.ac.uk/burks/foldoc/83/99.htm
[9] http://www-formal.stanford.edu/jmc/ascribing.html
[10] http://www-formal.stanford.edu/jmc/elaboration.html
[11] http://www-formal.stanford.edu/jmc/mcchay69.html
[12] http://www-formal.stanford.edu/leora/cs/

[Sorensen, 1997] Sorensen, R. (1997). Vagueness[13]. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab at the Center for the Study of Language and Information, Stanford University, Stanford, CA.

[Tarski et al., 1953] Tarski, A., Mostowski, A., and Robinson, R. M. (1953). Undecidable theories. In *Studies in logic and the foundations of mathematics*. North-Holland Pub. Co., Amsterdam.

[Tye, 2000] Tye, M. (2000). Vagueness[14]. In *Routledge Encyclopedia of Philosophy*.

[Williamson, 1994] Williamson, T. (1994). *Vagueness*. Routledge.

[13] http://plato.stanford.edu/entries/vagueness/

[14] http://www.rep.routledge.com/philosophy/