Towards a Comprehensive Standard Model of Human-like Minds

Thomas R. Hinrichs and Kenneth D. Forbus

Department of EECS, Northwestern University, 2133 Sheridan Rd. Evanston, IL 60208
{t-hinrichs, Forbus}@northwesetern.edu@aaai.org

Abstract

The Standard Model of particle physics has been an effective framework for describing entities of the domain relative to each other and thereby predicting missing particles. In this sense, it plays a similar role to the periodic table of the elements. Could something analogous work for AI? In this paper, we propose a dozen elements of a standard model of the mind and characterize their interrelationships. Naturally, this set reflects our own experience and biases and is not exhaustive. We also describe some of the ways in which this approach to a standard model differs from that in physics and suggest some limitations.

Introduction

A standardized, quasi-universal model of the mind is appealing because it could provide a way to characterize the scope of a particular contribution, where it fits with respect to other efforts, and whether it extends, elaborates, or contradicts existing models, which motivated the development of a proposal by Laird et al. (in press). We agree that without a unifying framework, AI research can look like five blind men describing an elephant. On the other hand, an overly precise or premature model could close off or marginalize otherwise promising approaches. Moreover, the analogy to the Standard Model in particle physics breaks down in the following ways:

1. There are no obvious symmetry laws that will make it predictive.
2. There may be little agreement on which elements to include.
3. There is no conventional granularity of modeling and/or level of abstraction with which to enumerate and describe elements.

Consequently, the nature of what a standard model of the mind might look like is still an open issue. The Laird et al. proposal comes from decades of experience with cognitive architectures that started out focusing on skill learning. Following Newell (1990), their standard model takes the form of a cognitive architecture, albeit a very abstract generalization that can describe systems based on both production rules and graph-based computations. Our focus is different. Just as the software running on today's computer architectures is only weakly specified by the hardware architecture, much of what is interesting about human-like minds appears to be at the level of their software, i.e. the kinds of knowledge, reasoning, and learning that they can do. For example, it would not be surprising to find that many other mammals also exhibit the power law of learning[1]. In that case, how much is the proposed standard model telling us about human-like minds, as opposed to minds more broadly? Given that, as best as we can tell, a signature feature of human-like minds is extraordinarily rich conceptual structures, any standard model of minds needs to focus on contents as well as on architecture. Perhaps a map might be a better metaphor, for some aspects of understanding minds.

With these caveats in mind, in this paper we suggest a first cut at elements of a standard model that are as much concerned with contents as processing. It will obviously be very schematic, and biased towards areas where we have worked previously. But we think this different perspective may help broaden the conversation and help lead towards a better framework.

A Generalized Model

We begin by specifying our assumptions, what constitutes an element of the model, and how elements are related. When we talk about a human-like mind, we assume some high-level criteria, such as *agency*, *adaptability* over diverse time scales, and *bounded rationality*. Such a system has purposes, it changes in response to its environment and experi-

[1] We know of little direct evidence on this point, although there are arguments that rats can learn rules (Murphy et al. 2008) and learning spatial knowledge can be modeled by a power law (Yadav et al. 2010).

ence, and its behavior degrades gracefully with limited resources. These assumptions motivate and constrain the elements of a model.

Proposed Elements

We propose a dozen elements of a standard model that could serve as points of comparison between architectures. Any given agent architecture may include a subset or superset, but these serve to define some dimensions along which alternative concrete models can be compared.

Goal-Seeking Behavior

The hallmark of any mind, much less a human-like mind, is that it behave with purposes. The idea that a machine can do this is generally attributed to Weiner (Rosenblueth et al. 1943). By itself, goal seeking may be little more than a feedback mechanism, but when an agent's goals pertain to its own internal states or another agent's mental state, that requires addressing the problems of adaptability and bounded rationality.

Our particular model of goal-seeking behavior in the Companions architecture (Forbus and Hinrichs, in press) involves explicitly represented goals, goal tradeoffs, and activation levels. These activation levels can, in turn, be represented as fluent quantities in qualitative process models, allowing reflective deliberation to influence relative goal activation, and for goal activation levels to conditionalize other processes (Hinrichs and Forbus, 2016).

Analogy

We view analogy as a primitive component of minds. We do not treat analogy itself as a kind of inference because it is an assessment of similarity over *descriptions*, rather than *assertions* and therefore serves a role closer to unification. In our model, analogy is a building block on which inferences and other operations are often built. Analogical projection, inductive generalization, memory retrieval, semantic disambiguation, visual classification, and even theory of mind are all based, in our modeling, on a foundation of analogy.

There are many alternative models of analogy itself, all centered on a notion of similarity, but our particular model is based on structure mapping, as implemented in the Structure Mapping Engine (Forbus et al. 2016).

Learning

Machine learning is, of course, a major sub-discipline of AI and reflects the importance of adaptation in any model of a mind. A salient distinction to make is between *skill learning* and *concept learning.* A standard model should account for both, regardless of learning mechanism(s).

A central part of our model of learning is analogical generalization, as embodied in SAGE (McLure et al. 2015). SAGE incrementally learns generalized concepts through merging sufficiently similar examples via analogy. SAGE has been applied to word sense disambiguation (Barbella & Forbus, 2013) and spatial concept learning from sketched depictions (McLure et al. 2015).

SAGE generalizations plus qualitative process representations have also been used to account for conceptual change phenomena (Friedman, 2012), and in combination with multimodal interaction it has been used to model vocabulary acquisition through representational change (Kandaswamy, 2016).

In addition to analogy, learning relates to other elements of a standard model. It can be thought of as memory plus inference. With the right inductive bias, an agent can learn a qualitative model to help performance on some planning tasks (Hinrichs & Forbus 2012).

Memory

We see memory, especially episodic or autobiographical memory, as a crucial element of a mind. Memory is the raw material of learning, but it also provides the historical context that serves as a viewpoint, or a sense of self.

Efficiently retrieving episodes or cases from an associative memory quickly runs into the problem of bounded rationality. As experience accumulates, the ability to find salient precedents or explanations should not degrade noticeably. Our answer to this problem is a model of long-term memory retrieval called MAC/FAC (Forbus, Gentner and Law, 1995). MAC/FAC is a two-stage model of retrieval in which an inexpensive feature-based filter feeds candidate cases to the analogical second stage. The FAC stage uses analogy to return the most similar case.

Inference

Inference suggests the ability to transcend simple stimulus-response and engage in deliberation. Primitive inference covers deduction, induction and abduction, but also analogy and Bayesian inference. Structured, multi-step inferences include planning and And-Or tree solving, for example. Minds are not solely recognition machines, but are capable of synthesis and indirect reasoning. Although bounded rationality would appear to rule out exhaustive theorem proving, some amount of inference is a necessary element of any standard model.

Representation

Human-like minds support inference, communication, naming and reflection. These capabilities demand rich, structured representations. A standard model need not commit to how much representation is symbolic vs. sub-symbolic, but there must be some way to drive linguistic expression and understanding and reasoning about objects not currently in an agent's perceptual field of view. Moreover, because human-like minds know a great deal about the world, knowledge is interconnected in ontologies, rather than solely limited to task-specific policies or recognition networks. We do believe that there is overwhelming evidence at this point for the necessity of relational representations in

human cognition, although whether or not symbolic representations plus statistics are sufficient for human cognition, versus distributed representations also playing a role, is still an open question at this point (Forbus et al. 2017).

Theory of Mind and Social Interaction

The recognition of, accommodation for, and interaction with other minds occurs remarkably early in human development. Children learn rapidly via imitation, and human children seem unique in their urge to be helpful to others. Given the rise of software assistants and robots that need to work with people, considerable efforts have been made recently on modeling theory of mind and social interaction. Qualitative reasoning can be used for modeling blame assignment for event outcomes (Tomai & Forbus, 2007), and analogy has been used to model moral decision making (Dehghani et al. 2008) and how children learn in a false belief training study (Rabkina et al. 2017).

To us, social interaction builds on a theory of mind and covers all kinds of communication, most notably verbal language. Vygotsky (1962) and Tomasello (1999) both argue for the centrality of social behavior as a means of bootstrapping intelligence. While one might quibble about whether theory of mind and social interaction should be lumped together or treated as distinct elements, there is little doubt they should be part of a standard model.

Attention

Bello and Bridewell (in press) convincingly argue that attention should be a first-class element of any model of agency. They point out that any interesting form of agency involves choice, and under bounded rationality this must entail control or focusing of attention.

We might further suggest that attention is modulated by emotion (i.e., arousal levels), and in turn influences what is remembered, serving as a kind of lossy compression for percepts.

Emotion

Emotion is increasingly appreciated as a central part of cognition (Gratch & Marsella, 2004; Minsky, 2006). In Appraisal Theory, beliefs about a situation are evaluated by calculating appraisal variables (e.g. desirability, likelihood), from which emotions about those specific entities and beliefs are generated. These emotions trigger coping strategies for dealing with the situation. Coping strategies can include internal strategies, like giving up on a cherished goal, or external, such as working when one would rather be playing when a deadline looms. A standard model of human-like minds should include emotion as a core element.

One treatment of emotion as it affects problem-solving behavior is described in (Wilson et al. 2013), where analogical retrieval is used for an initial appraisal, modeling the rapid response that organisms seem to have. A later cognitive appraisal helps balance the initial response, and retrospective analysis before consolidation of the problem-solving episode into long-term memory provides a more balanced view of what happened. Thus operations which initially were avoided (a coping strategy) because they seemed too hard but eventually were found to lead to solving a hard problem were stored as more desirable, thereby changing future behavior.

Reflection

Reflection is the ability to think about one's thinking. The need for reflection in a standard model derives from the constraint of bounded rationality. A purely reactive system is bounded, to be sure, but lacks the open-ended reasoning abilities that are a hallmark of human-like minds. Reflection need not entail a distinct level with different representations, but can simply be a mechanism for associating names with distinguishable mental states on demand. Reflection is the means by which an agent can not merely do a task or be in a state, but know what it is doing and both reason about, and communicate about, its internal state.

Causal and Qualitative modeling

Models provide ways of reasoning abstractly and indirectly, allowing efficient projection and explanation. Modeling, especially qualitative modeling, is relevant to a standard model because it is applicable to modeling any continuous system, that is, social relationships (e.g. degree of friendship) and mental state properties (e.g. degree of difficulty of a problem), in addition to the physical world. Qualitative representations allow reflection and reasoning about states defined with respect to relations between quantities, without requiring them to be pre-enumerated or named.

Spatial and Temporal reasoning and context

Clearly, human-like minds must reason about time and space. What makes this worth including in a standard model is the indexical context they provide in the form of Here and Now. It is hard to imagine how an agent could be an individual without some locality in time and/or space. Even without physical embodiment, the ability to reason about that context is critical for reasoning about other agents and *their* ego-centric worldview.

Omitted elements

Our list is almost certainly incomplete, largely because it focuses on areas in which we have worked. Some obvious omissions include:

1) Physical embodiment. Our group is starting to look at vision and gesture recognition via a Microsoft Kinect™ which may lead to an increased emphasis on embodiment.
2) Real-time behavior. We focus primarily on what Newell (1990) calls the social band of cognition, so this has not been a major issue for us, though bounded rationality certainly is.
3) Ethics. There has been a great deal of hand-wringing lately about building in ethics from the start. However,

given that people aren't born with innate ethics, we believe that our approach of using analogy based on stories in MoralDM provides a better approach. It is also important for artificial minds to have a sense of *empathy*, which likely relies on analogy between self and others.

Interrelationships

As elements, these are coarse building blocks. How do they fit together? What dependencies exist between them? Is there a common substrate or can they be treated as stove-piped systems? A set of elements is not a model until their interrelationships are constrained. Table 1 is an attempt to present these interrelationships in the form of pairwise compositions. It helps to think of the model elements as abstract operations on mental states, rather than as dynamic processes or static structures. Thus, "representation" is the mapping from mental states to symbolic names, rather than any particular encoding. Because the composition is not symmetric, entries in the cells should be taken to denote a mental state or process that results from or is supported by the application of the row heading element to the column heading element. So, for example, inductive model acquisition results from the application of learning to modeling, whereas modeling of the learning process can support the design of experiments. To the extent that element compositions have well-defined outcomes, some table of this sort, while not predictive, may suggest the breadth of phenomena covered (or missing) by a model.

Unfortunately, lack of space makes for some extremely abbreviated entries. For example, the cell for emotions applied to spatial & temporal modeling simply says: "Appraisals w.r.t. res. limits" which conceivably might fail to evoke the sense of panic that can result from the realization of rapidly diminishing time before a deadline.

Although this table is a first pass and has some holes in it, it does suggest that there could be value in relating proposed modeling elements. A more complete model would better cover the space of behaviors and capabilities we see in human minds, and missing or unclear relationships between elements may indicate elements at the wrong level of abstraction or of the wrong kind. Some of the patterns we do see in this model lead us to three conjectures:

Conjecture 1: Every plausible reasoning mechanism can be reduced to some combination of similarity assessments which we would call analogy. In other words, analogy is ontologically prior to induction and abduction (and for the most part, the induction of rules precedes the formal deductive application of those rules)

Conjecture 2: Every lossy compression mechanism involves a similarity judgement of some sort. We have suggested above that a critical ingredient of bounded rationality can be modeled as a kind of compression. We conceive of cognitive compression as the elision of "similar" or non-informative content. SAGE is our model of how this is done.

Conjecture 3: Qualitative state representations reduce or condense descriptions to their essences, supporting inference amid the constraints of bounded rationality. For example, one way to look at qualitative process models is as a highly condensed implicit representation of state machines, in which state transitions are inferred from quantity relations, rather than explicitly enumerated. This makes qualitative representations effective not merely for physical processes, but also for representing internal mental states and continuous aspects of social reasoning (Forbus & Kuehne, 2005).

Although such conjectures are not direct consequences of the model in the way that missing particles are predicted by the physics Standard Model, they are an informal result of thinking about the big picture in a way that is facilitated by having a more comprehensive model.

Future Prospects

The effort of trying to achieve some consensus on a standard model of the mind is worthwhile, regardless of whether it produces a single standard model. By identifying the dimensions along which models differ, it will undoubtedly encourage researchers to think about where the holes are in their accounts, and islands of agreement may lead to integrated models that encompass more diverse phenomena, rather than fine tuning isolated mechanisms. We further suggest that it is important to think of such a model as not just an architecture, but with additional commitments to the kinds of knowledge and reasoning that must be supported. In other words, a map as well as an architecture, whose interconnections might be used to fill in missing pieces, just as was done with the Periodic Table in chemistry.

Finally, there are many kinds of minds in nature, so speaking of "the mind" seems unduly limiting. Building models that are general enough to compare and contrast across species would sharpen what we mean by human-like minds as well as enable us to understand intelligence more deeply.

Acknowledgements

This research was supported by the Air Force Office of Scientific Research and the Machine Learning, Reasoning, and Intelligence Program of the Office of Naval Research.

	Goals	Analogy	Learning	Memory	Infer.	Repr.	Modeling	Refl.	Emot.	T.O.M.	Space / Time	Attention
Goals										Comm. goals	Factored planning	Goal sacrifice
Analogy	Deontic reasoning						Conceptual change			Project. self to other	Spatial & temporal analogies	
Learning	Explicit learning goals	Analogical generaliza-tion		Reinforce-ment learning			Inductive model acquisition	Learning strategy selection	Learning coping strate-gies	Learn-ing from others	Inducing causal propensity	Learning salient predictors
Memory	Retrieving appropriate goals	Analogical retrieval			Recon-structive memory						Priming	Case encoding
Inference	Planning	Analogical projection		Case-Based Reasoning								
Repr.	Goal ontology	Analogy ontology	Reprs. of multistrategy learning								Qualitative S-T reprs.	
Models	Subgoal & tradeoff repns.	ID model-ing as-sumptions	Experiment design	Adopting known models	Envision-ment & explana-tion							
Refl.	Goal rela-tions & activation	Reasoning about mappings	Evaluating effectiveness of methods	Relations between historical contexts	Progress estimation	Deliberate re-rep.	Optimizing flows		Attrib. beh. to emotion			Focusing attention
Emot.	Appraisals w.r.t. goal relns.	Surprise response	Curiosity as a driver	Appraisals w.r.t. past	Coping strategies	Biasing repres. choices					Appraisals w.r.t. res. limits	
T.O.M.	Imputing goals to others	Imitation	Instructing others	Attributing beliefs to other's experience	Predicting & ex-plaining behavior	Detecting ontological mismatch	Grounded comm. via shared models	Reflecting assessment back to self	Empathy		Transform to other's P.O.V.	Attributing miscues to attention limits
Space / Time	Activation from space-time intersection	Context-sensitive analogies	Choosing learning strat via temporal budget	Estimating from precedents	Inferring spatial & temporal relns		Process boundary conditions					
Attention	Tracking	Analogical salience assessment	Filtering inputs to learning						Arousal levels			

Table 1: Interrelationships between elements`

References

Barbella, D. and Forbus, K. 2011. Analogical Dialogue Acts: Supporting Learning by Reading Analogies in Instructional Texts. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence* (AAAI 2011), San Francisco, CA.

Bello, P. and Bridewell, W. (In press) There Is No Agency Without Attention. *AI Magazine*.

Chen, K. and Forbus, K. D. 2017. Action Recognition from Skeleton Data via Analogical Generalization. *Proceedings of the 30th International Workshop on Qualitative Reasoning*, Melbourne, Australia.

Dehghani, M., Tomai, E., Forbus, K., Klenk, M. (2008). An Integrated Reasoning Approach to Moral Decision-Making. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (AAAI)*. Chicago, IL.

Forbus, K. D., and Hinrichs, T. (In press) Analogy and Qualitative Representations in the Companion Cognitive Architecture. *AI Magazine.*

Forbus, K. D., Ferguson, R. W., Lovett, A., and Gentner, D. 2016. Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, DOI: 10.1111/cogs.12377, pp 1-50.

Forbus, K., Gentner, D. and Law, K. 1995. MAC/FAC: A model of Similarity-based Retrieval. *Cognitive Science*, 19(2), April-June, pp 141–205.

Forbus, K, Klenk, M. and Hinrichs, T. 2009. Companion Cognitive Systems: Design Goals and Lessons Learned So Far. *IEEE Intelligent Systems*, 24(4), 36-46.

Forbus, K. D. and Kuehne, S. (2005). Towards a qualitative model of everyday political reasoning. *Proceedings of the 19th International Qualitative Reasoning Workshop*, Graz, Austria, May

Forbus, K., Liang, C. & Rabkina, I. (2017) Representation and Computation in Cognitive Models. *Topics in Cognitive Science*, 9:694-718, DOI: 10.1111/tops.12277.

Friedman, S. E. 2012. Computational Conceptual Change: An Explanation-Based Approach. Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois.

Gratch, J., & Marsella, S. 2004. A Domain-independent Framework for Modeling Emotion, *Journal of Cognitive Systems Research*, 5(4), 269–306.

Hinrichs, T. & Forbus, K. 2012. Learning Qualitative Models by Demonstration. *Proceedings of AAAI-2012*.

Hinrichs, T. and Forbus, K. 2016. Qualitative Models for Strategic Planning. *Advances in Cognitive Systems*, Volume 4, 2016, pages 75-92.

Kandaswamy, S. 2016. Comparison driven representational change. Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois.

Laird, J. E., Lebiere, C. & Rosenbloom, P. S. (In press). A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine*.

McLure, M.D., Friedman, S.E., & Forbus, K.D. 2015. Extending analogical generalization with near-misses. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX*.

Minsky, M. 2006. *The Emotion Machine*. Simon & Schuster, New York.

Newell, A. (1990) *Unified Theories of Cognition*. Harvard University Press.

Rabkina, I., McFate, C.,Forbus, K.D., and Hoyos, C. 2017. Towards a Computational Analogical Theory of Mind. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, England, July.

Rosenblueth, A., Wiener, N. & Bigelow, J. 1943. Behavior, Purpose, and Teleology. *Philosophy of Science*, 10(1), 18-24.

Tomai, E. and Forbus, K. (2007). Plenty of Blame to Go Around: A Qualitative Approach to Attribution of Moral Responsibility. *Proceedings of Qualitative Reasoning Workshop 2007*, Aberystwyth, U.K.

Tomasello, M. 2001. *The Cultural Origins of Human Cognition*. Harvard University Press.

Vygotsky, L. 1962. *Thought and Language*, MIT Press, 1962.

Wilson, J., Forbus, K., and McClure, M. 2013. Am I Really Scared? A Multi-phase Computational Model of Emotions. *Proceedings of the 2nd Conference on Advances in Cognitive Systems*.