A Standard Model of Mind:
AAAI Technical Report FS-17-05

Anatomy of a Task: Towards a Tentative Taxonomy of the Mind

David Sekora,[1] **Samuel Barham,**[1] **Justin Brody,**[2] **Don Perlis**[1]
[1]Department of Computer Science, University of Maryland - College Park
[2]Department of Computer Science, Goucher College

Abstract

A standard model of mind will involve not only an architecture but also a set of capabilities. Ideally, the two should inform one another at a deep level, as an architecture is what both enables and constrains capabilities. In that spirit, we consider in some detail a routine and (deceptively) simple robotic task. From it, we build out a substantial list of capabilities that appear essential for a general-purpose execution of the task. We argue that this type of exercise is an indispensable step toward the establishment of a baseline for the comparison of cognitive architectures, and that the resulting taxonomy can inform the synthesis of a standard model of the mind.

Introduction

The idea that the cognitive sciences might have developed sufficiently, at long last, to justifiably produce a *unified theory of cognition* — i.e., one that describes a "single set of mechanisms for all of cognitive behavior" — is not new at all; it can be traced back at least to (Newell 1990). The notion, on the other hand, that there might be consensus enough among cognitive scientists to begin to compile a *standard model of mind* is far more surprising, and it is just this potential for consensus that forms the basis for the forthcoming (Laird, Lebiere, and Rosenbloom 2017). We jump into the discussion with our own ideas about what sorts of abilities a standard model of mind must subsume in order to be truly general and human-like, as well as how one might use this as a framework to assess cognitive architectures.

We focus primarily on a story about a robot and a task she has been given — a task that turns out to be more complicated than it at first appears. The story is annotated with the capabilities an intelligent agent would need to accomplish the feats therein. After the story, we collect these capabilities into a unified list, which we believe can serve as the basis for a general list of desiderata for intelligent agents (and hence their cognitive architectures).

Finally, we briefly discuss how one might evaluate existing and future cognitive architectures in light of our list, using our own ALMA system as an example.

Currently, we are unaware of any system that possesses all of these capabilities to any advanced degree, nor of any system that could successfully navigate a scenario like the one we here lay out; we don't believe this to be a coincidence.

Robbie — A Day in the Life

This story is inspired by a similar story featured in (Perlis et al. 2013), but here we flesh it out considerably.

We use a bracket notation to show where items from our list of desiderata (presented in a later section) appear in the story. For example, the marker [1a] would mean that the preceding sentence features perception of the outside world.

> *Robbie the robot is in trouble. She has been tasked with retrieving a book for us from room 128, but everything has gone wrong!*

Already, we're hinting at one of our major themes — in the real world, things often go wrong, and an intelligent agent will need to be able to cope. This functionality is commonly referred to as *perturbation tolerance*.

> *We made our request at 11:30AM, asking that Robbie bring us the book by noon. Robbie ran a simulation[3a] and determined that, given the distance she had to cover to reach room 128 and the motions involved in picking up and carrying a book, this was well within her capabilities[4e]. So, she planned out a sufficient route and set off[1e,3b]. Once her initial planning for the current task was done, she allocated the necessary resources to navigate and watch for anomalies, opting to use the rest of her processing power to continue working on a math problem we interrupted her in the middle of solving[3h].*

Here, we see that Robbie has some interesting cognitive features. She can conduct physics simulations and make multi-step plans. Implicitly, she must have been able to break up a complex task into sub-parts ("retrieve the book" becomes "go to the book," "pick up the book," and "return here while still holding the book"). She has an understanding of her own capabilities and limitations. Subtly, she stops at a "sufficient route" to complete the task within the time limit, rather than continuing to search until she has found the "optimal route" — this requires her to understand that the actions of reasoning and planning themselves take time!

Finally, the last sentence makes it clear that Robbie is more than just an object retrieval bot. As a persistent, gen-

eral agent, she has more to do and think about than just an immediate command.

The trouble starts about halfway through the journey. One of the hallways Robbie had planned to use was closed for maintenance! Robbie realizes that she will not be able to complete her original plan, and computes a new, slightly longer route using a different hallway[3b]. She determines that she will still be able to finish in time, and so sets off again.

Now Robbie has run into her first obstacle. She could hardly be called an "intelligent agent" if she broke down or quit when the first problem came up. She has the ability to learn new information, incorporate it into her database, identify that this will prevent her from completing her current plan, and make a new plan, all while keeping her original goal in mind.

Robbie finally reaches room 128 — or, to be precise, the room marked on her internal map as room 128[5c]. The problem? Before opening the door, she notices that the room number says 123, not 128[1a,4d]! Here, a contradiction arises: the room in front of her appears to be both room 123 and 128, but she knows that a room in this building can only have one number. A lesser agent might give up, but luckily Robbie has strategies for dealing with apparent contradictions.

For an intelligent agent to *deal with* problems, the agent must first be able to *detect* those problems. This particular example requires some degree of visual processing, but any such problem will at a minimum require the agent to have a set of expectations, a means of checking those expectations against reality, and strategies for resolving any conflicts.

Robbie reasons that either the room in front of her is room 123, the room in front of her is room 128, or one of her base beliefs about how room numbers work is flawed[2h]. Robbie has a record of her past reasoning[2j], and so she knows that she has only limited evidence for the room being either 123 or 128. Therefore, she decides to trust her base beliefs for the moment, and directs her efforts towards obtaining more evidence[3b] for the two more likely possibilities.

Here we see that Robbie is able to identify a plausible set of beliefs that could be mistaken, and furthermore able to leverage her memory to estimate how reliable each belief is. Once she narrows it down to two suspicious beliefs, she applies a common contradiction resolution strategy: gather more data!

After some thought, Robbie finds a potentially relevant fact in her knowledge base: "Room numbers appear in sequence."[2f]. This looks promising; there may be an even better method for resolving this situation buried deep in her knowledge base somewhere, but she has a strict time constraint and so cannot afford to be picky[4a]. She quickly makes and executes subplans to read the room numbers of the adjacent rooms, and finds that the rooms on either side are numbered 126 and 130. This evidence supports the hypothesis that the mystery room is in fact room 128. As she approaches the room once more, she reflects on how it might be that she saw 123 instead of 128[3d]. She does some quick simulations[3a] and realizes that, if you rub away the left part of an 8, it looks like a 3. She makes note of this fact for future room-identification scenarios[2g].

An intelligent agent needs to be able to identify new goals and make plans on the fly. Time continues to be an important consideration — as soon as a viable plan that doesn't take too much time is found, Robbie springs into action. Once the plan is executed, simulation capabilities are once again highlighted, along with some nominal curiosity and learning ability.

Robbie still isn't 100% sure that the room is room 128, but now she has much more evidence for it, and time is of the essence[4a]. So, she pushes on the door — but it won't budge! She consults her knowledge base, but — alas — to no avail: this door has a handle on it, and Robbie hasn't learned anything about handles yet[4e]!

"General purpose cognitive agent" does not, of course, mean "omniscient robot." There will be many times when a cognitive agent (no matter how general) simply does not understand its environment, or its interactions with it. In such cases, the agent must have other methods of coping. Robbie must be able to ask for advice and learn from it.

So Robbie does what anyone in over their head should do — she asks for help[3c]! Robbie phones her researcher and briefly explains the situation:

ROBBIE: I am requesting assistance. I attempted to open the door using my strategy of "push on the surface of the door," but this did not work. I am sending you a picture of the door[1b,1c].
RESEARCHER: Ah, yes. I see what's happened — this type of door has a handle, so you're going to have to use your arm to turn it before you can open the door. Here, I'll send you a video of a person opening a door. Hold on just a second. When you get it, watch it and try to learn from it[1d,2g].

While it may be possible that communication skills are not strictly necessary for an agent to qualify as "intelligent," the only intelligent agents we currently know about (humans) have sophisticated forms of communication, and the two concepts certainly give the appearance of being inextricably linked. In any case, for an intelligent agent to be able to competently interact with humans, it will need at least a rudimentary understanding of language — both how to interpret it, and how to produce it.

The process by which Robbie learns how to open the door may need to be a bit more involved than simply being told or watching a video. That would be ideal, since humans are able to do these types of advice-based and single-example learning quite well, but the current trend in AI research is more in the realm of training on hundreds or thousands of examples (which is of course also very important — humans also make good use of this type of learning). At any rate, we aren't particularly concerned with the specifics here — the

important thing is that an intelligent agent must have some mechanisms in place for learning new facts and skills.

Robbie successfully opens the door, and files this newfound skill away for future use[2g]. She scans the interior, and finds to her dismay (or whatever it is that robots feel when their expectations aren't met) that it's a mess! The floor is strewn with hundreds of books; perhaps there was an earthquake, or maybe it was just those crazy graduate students again. She does some quick math to figure out how long it will take her to sort through the mess to find the requested book[3a], and determines it to be highly unlikely that she should find it in time[4a]. Robbie realizes things are not going well, and so phones her researcher once more for advice[3c,4f]. This time, he agrees that it's likely impossible for her to get the book in time, and tells her to call the whole thing off. Robbie obliges, and heads back to the lab.

A perhaps underrated strategy for dealing with problems is to know when to give up. This will involve at least two abilities. The first is the ability to dynamically prioritize live tasks on the basis of a cost-benefit analysis; the second, implied by the first, is the ability to predict one's probability of success (or lack thereof) in a given task-instance.

Back in the lab, Robbie doesn't just shut off, even though she no longer has a task. Instead, she returns to working on that math problem[3h].

A persistent agent must be, well, persistent.

This is for the AI researcher likely a satisfying ending to the story. However, consider briefly an *alternative* ending, which illustrates yet another important cognitive ability — namely, the ability to imagine novel goals, or to infer goals that other agents are likely to possess; and, in response, to invent, prioritize, and initiate novel tasks. Watch:

[...] *Robbie obliges, and starts to head back to the lab. However, as she approaches the door, a subprocess realizes that it would be helpful in the future for the room to be in order, should she or someone else be tasked with finding a book again[3d,3f]. She infers that this goal also seems desirable to her researcher[3e], and so assigns herself the task of organizing the books in the room, marking it as the active goal but noting that it can be interrupted if she is requested elsewhere. The only other source of intrigue at the moment is that math problem from before, but it has relatively lower priority[3g]. So, she gets to work cleaning the room with all the alacrity characteristic of her programming.*

Intelligent agents need to be able to plan for long-term benefit. Plus, a robot might as well make itself useful!

List of Desiderata for Intelligent Agents

We now present a list of capabilities, primarily derived from the story, which a general-purpose intelligent agent might reasonably be expected to possess. We do not claim this to be a complete list, only that a "smart" robot would likely need most of these. For convenience, we have grouped the list items into rough categories, although the category headings are meant more as guidelines than rigid laws.

So, without further ado, we believe it safe to say that an intelligent agent should be able to:

1. **Interaction**:
   - (a) Perceive the outside world;
   - (b) Bring attention to spatial entities (e.g. pointing);
   - (c) Generate simple language;
   - (d) Understand simple language;
   - (e) Move self and other objects;
2. **Knowledge and Learning**
   - (a) Identify objects;
   - (b) Keep track of real-valued quantities (such as counts);
   - (c) Learn new objects and how they behave (individuals and classes);
   - (d) Deliberately affect its perceptions (e.g. move to get a better viewing angle);
   - (e) Track own actions and processing in real time (e.g. efference copy, as in (Brody, Perlis, and Shamwell 2015));
   - (f) Maintain a knowledge base (KB);
   - (g) Update the KB with new information ("learning");
   - (h) Make inferences based on the KB;
   - (i) Maintain information about others' knowledge;
   - (j) Keep a detailed history of own activity and perception;
3. **Goals, Planning, and Acting**
   - (a) Simulate behavior in imagination (for use in vision/planning);
   - (b) Make and execute plans to achieve goals, including backup plans where appropriate;
   - (c) Ask for help effectively (knowing whom and how to ask);
   - (d) Identify new goals, including ones for future or long-term benefit;
   - (e) Identify needs of others;
   - (f) Be helpful (as appropriate);
   - (g) Keep track of priorities and rearrange them as necessary;
   - (h) Seek knowledge as a general goal, when consistent with other goals;
   - (i) Identify overly complex plans, and have strategies for dealing with them (prune, get help, give up);
4. **Real-World Considerations**
   - (a) Control activities (including inference) to respect real-time constraints;
   - (b) Forget and relearn when necessary;
   - (c) Possess and apply contextual awareness;
   - (d) Detect anomalies in the world and in reasoning, and have strategies for dealing with them;
   - (e) Know its own capabilities and shortcomings;

(f) "Take stock": how are things going overall in the short/medium/long term?;

5. **Special Category Distinctions**

(a) Distinguish self from other;

(b) Distinguish parts from wholes;

(c) Distinguish appearance and thought from reality.

If you read the story carefully, you may notice that not all of these capabilities are directly referenced. For the most part, these are additional capabilities such as [4c] "Possess and apply contextual awareness" which underly much of the thought and action taking place throughout, even though they aren't the central focus of any single paragraph.

Notable Omissions

It may seem that some obvious things are missing from this list. Take, for example, "communication between submodules." Minds are complex systems with lots of specialized parts, and to navigate the world successfully an agent will almost certainly need to be able to use several of them together at once.

However, this is the advantage of our approach. Perhaps submodules need to be able to communicate, or maybe specialized submodules aren't even required at all — it makes no difference here! These are architectural considerations, whereas we (at least in this paper) are concerned purely with capabilities.

Of course, we still have likely omitted something that should be on the list. Again, this list is not meant to be complete or final — to the contrary, we encourage discussion and iteration on both the content and structure of the list. In our view, the development of such a list as this represents a significant step on the path to a standard model of mind, and can inform future work on intelligent agents and cognitive architectures.

Existing Systems and Architectures

Of course, there are many systems and architectures that have been thought up and implemented prior to the creation of this list of desiderata. Originally, we had intended to evaluate a number established and/or recent systems, architectures, and models (such as SOAR (Laird 2012), ACT-R (Anderson et al. 2004), MIDCA (Cox 2013), DIARC (Schermerhorn et al. 2006), and HoA (Chaouche et al. 2014)) with respect to our list. However, realistic cognitive architectures like the ones we mentioned are tremendously complex, so a "complete" analysis of each one would draw too much focus from the rest of the paper, and anything less would be doing the creators of these systems a disservice. Instead, we invite these groups to examine their own work in the context of our list. To demonstrate how this might be done, we will take a look at our own ALMA (Purang 2001) system.

Case Study: ALMA

ALMA (short for **A**ctive **L**ogic **MA**chine) is a general-purpose reasoner which implements the titular active logic formalism. Active logic is a form of first-order logic derived from step-logic (Elgot-Drapkin, Miller, and Perlis 1991), which is built to accomodate reasoning situated in time.

1. Interaction ALMA is a reasoning system, not a full embodied agent, so many of the items in this section don't really apply. However, systems have been built (e.g. (Josyula, Anderson, and Perlis 2004)), and are being built in our current work, that incorporate ALMA as a core component which are capable of all of these.

2. Knowledge and Learning Being a reasoning system, these items are the bread and butter of ALMA. [2f - 2h] are central components of almost any reasoning system, and [2e] and [2j] are central features that distinguish ALMA from other systems. [2a], [2c], and [2i] are again more in the domain of a larger, integrated system. [2b] is a current research focus for the core of ALMA, and [2d] is an important piece of a demo currently under construction in which ALMA is being used as the control center for a robot.

3. Goals, Planning, and Acting ALMA has very rudimentary planning facilities like [3b], specifically for executing tasks at future times, and [3a] and [3g] are active research topics. There is still much to be desired on the items in this section, although again this is mitigated by the fact that it is meant to be combined with other systems (which would handle planning and execution).

4. Real-World Considerations [4a] is the defining feature of ALMA. [4b] and [4d] have been written about extensively in papers relating to active logic (Miller and Perlis 1993), and this work has been partially implemented. There has been some work on [4e] (Nirkhe et al. 1997), but it has not been implemented. Not much has been done about [4f].

5. Special Category Distinctions [5a] is arguably covered — an instance of ALMA is aware of its own existence as an entity, although this hasn't been used for very much. Much has been written about how to handle [5c] in papers such as (Miller and Perlis 1993), although implementation has lagged behind here.

Verdict Many of the items on our list of desiderata are represented or at least considered; this is perhaps to be expected, as the system and its underlying logic have been worked on by some of the very same minds as this paper.

Other Lists

Various other lists exist which can, at first blush, seem very similar to ours. Here, we'll pick out a couple of representatives to demonstrate how our list distinguishes itself.

Constraints on Mind

In 1980, Allen Newell briefly presented a set of 13 constraints on the human mind (Newell 1980), a set which was later expanded upon by Anderson and Lebiere (Anderson and Lebiere 2003). Our list mostly subsumes this one, barring the last few items (on the 2003 formulation): "acquire capabilities through development", "arise through

evolution", and "be realizable within the brain". This is because that list is about constraints on the *human* mind, whereas we are looking at the more general class of intelligent minds, and so are not concerned with how the mind arises or whether it matches up with a particular physical system.

Architecture

Orthogonal to papers such as (Laird, Lebiere, and Rosenbloom 2017), which center more on the cognitive system architecture (the "how"), we focus on the possible types of thought (the "what"). We intend this list to serve as a complement to the more technically-minded papers in this collection, as well as to future papers — in order to design an architecture, you first have to have some idea of what you want that architecture to be able to do.

Conclusion

We began by working through a scenario of our design and considering what cognitive capabilities would be necessary to successfully negotiate it. We then collated these capabilities, and extended and generalized the resulting list into its current form. This list is not meant to be a final, definitive list of all the capabilities an agent must have to be considered intelligent in a general sense. However, it can serve as the prototype for such a list; as such, we welcome discussion and modification.

If the research community is able to reach a consensus about what the contents of this list should be, then in turn it can serve as a foundation for a standard model of the mind.

References

Anderson, J. R., and Lebiere, C. 2003. The newell test for a theory of cognition. *Behavioral and brain Sciences* 26(5):587–601.

Anderson, J. R.; Bothell, D.; Byrne, M. D.; Douglas, S.; Lebiere, C.; and Qin, Y. 2004. An integrated theory of the mind. *Psychological Review* III(4):1036–1060.

Brody, J.; Perlis, D.; and Shamwell, J. 2015. Who's talking - efference copy and a robot's sense of agency. *AAAI 2015 Fall Symposium*.

Chaouche, A.-C.; Seghrouchni, A. E. F.; Illi, J.-M.; and Sadouni, D. E. 2014. A higher-order agent model, with contextual planning management for ambient systems. *LNCS Transactions on Computational Collective Intelligence* XVI:146169.

Cox, M. T. 2013. Midca: A metacognitive, integrated dual-cycle architecture for self-regulated autonomy. *UMIACS Technical Report No. UMIACS-TR-2013-03*.

Elgot-Drapkin, J.; Miller, M.; and Perlis, D. 1991. Memory, reason, and time: the step-logic approach. In Cummins, and Pollock., eds., *Philosophy and AI: Essays at the Interface*. Cambridge, MA: MIT Press.

Josyula, D. P.; Anderson, M. L.; and Perlis, D. 2004. Domain-independent reason-enhanced controller for task-oriented systems - director. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-04)*, 1014–1015.

Laird, J. E.; Lebiere, C.; and Rosenbloom, P. S. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*.

Laird, J. E. 2012. *The Soar Cognitive Architecture*. The MIT Press.

Miller, M., and Perlis, D. 1993. A view of one's past and other aspects of reasoned change in belief. *(dissertation)*.

Newell, A. 1980. Physical symbol systems. *Cognitive science* 4(2):135–183.

Newell, A. 1990. *Unified Theories of Cognition*. Cambridge, Massachussetts: Harvard University Press.

Nirkhe, M.; Kraus, S.; Miller, M.; and Perlis, D. 1997. How to (plan to) meet a deadline between now and then. *Journal of Logic and Computation* 7:109–109.

Perlis, D.; Cox, M. T.; Maynord, M.; McNany, E.; Paisner, M.; Shivashankar, V.; Hand, E.; Shamwell, J.; Oates, T.; Du, T.; Josyula, D.; and Caro, M. 2013. A broad vision for intelligent behavior: Perpetual real-world cognitive agents. *2013 Annual Conference on Advances in Cognitive Systems: Workshop on Metacognition in Situated Agents*.

Purang, K. 2001. Alma/carne: Implementation of a time-situated meta-reasoner. In *Proceedings of the International Conference on Tools with Artificial Intelligence*, 103–110.

Schermerhorn, P.; Kramer, J.; Middendorff, C.; and Scheutz, M. 2006. Diarc: A testbed for natural human-robot interaction. *AAAI*.