The 2016 AAAI Fall Symposium Series:
Artificial Intelligence for Human-Robot Interaction
Technical Report FS-16-01

Reasoning with Grounded Self-Symbols for Human-Robot Interaction

Justin Brody,[**] Samuel Barham,[*] Yue Dai,[*] Christopher Maxey,[*] Donald Perlis,[*] David Sekora,[*] Jared Shamwell[*]

[**]Goucher College; [*]University of Maryland College Park
justin.brody@goucher.edu

Abstract

We discuss Perry's notion of the *essential indexical* and the requirement that robots interacting with humans (and other robots) be able to reason about themselves in a grounded way. We describe an approach based on grounding symbols via an analogue of the neural mechanism of *efference copy* and approaching symbolic reasoning via *active logic* – a situated framework for logical and temporal reasoning.

The Essential Indexical

Robots that interact with humans should be able to make sense of indexical terms (such as "I" and "now"). For example, a robotic helper should ideally be able to interpret human utterances such as:

> *I am getting hungry. I don't want to eat now, but would like you to get me a snack from the kitchen in 10 minutes. Please make sure to check that you didn't spill anything after you're done, and make sure to clean yourself up too if needed.*

Understanding such a sentence will require, along with much else, an ability to use indexical notions of "I", "you", "now" and "in 10 minutes." A brief digression into philosophy will demonstrate that this is a more difficult problem than it first appears to be. See (Anderson and Perlis 2005) for a fuller account with explicit ties to robotics.

The meat of our digression comes from a seminal 1979 essay by Stanford philosopher John Perry. In it, he describes the peculiar experience of following a trail of sugar around a grocery store. Perry spent some time pushing his cart up the aisle on one side of a tall counter, then down the other, noting all the while that the trail grew thicker and thicker. Eventually the truth dawned on him: as he put it, "I was the shopper I was trying to catch." Perry spent the rest of his seminal paper trying precisely to characterize the changes in his beliefs that precipitated the change in his behavior -- which was, of course, that he stopped looking for the shopper with the torn sugar sack.

One should like to say that Perry came to believe that "I am the shopper making the mess." Things turn out not to be so simple. What Perry finds is that in a traditional Fregean framework of *de dicto* belief -- where belief is characterized as a particular relation between a subject and a proposition -- there is no room for indexicals like "I" or "now." For if we accept Perry's construal of the traditional "doctrine of propositions," then we agree that for a proposition S to equal proposition S' is precisely for S and S' to have not only the same truth-value, nor merely the same truth-conditions -- that is, merely share the same reference -- but that they have also the same intension, or sense. As Perry explains, "Atlanta is the capital of Georgia" and "Atlanta is the capital of the largest state East of the Mississippi" are not the same propositions -- though they are certainly true or false under the same conditions -- for I can clearly believe one but not the other without inconsistency, depending on what I believe about Georgia[1]. But then it cannot be that Perry came to believe "I am the shopper with the torn sack," because that sentence does not even identify a proposition. It is not true or false absolutely. It is ambiguous, depending on who says it. Nor does it help, as Perry points out, to argue that indexicals like "I" or "now" or "this" are communicative shortcuts standing in for some propositions α, β, or γ. For if we claim that all Perry came to believe was, in fact, "α is making a mess," where α unambiguously picks out Perry[2], we have not yet explained his subsequent change in behavior -- for we also believe this proposition, and yet haven't stopped reading to check our shopping carts. No; Perry must also have believed "I am α."

Perry goes on to describe in detail the way in which indexicals like "I" and "now" trouble the waters of tradition-

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] Hilary Putnam argued against fixing reference in general based solely on truth-conditional criteria. See (Lakoff 1990) for a good exposition.

[2] Perhaps, Perry suggests, α might be "the only bearded philosopher in a Safeway store West of the Mississippi."

al theories of belief, both *de dicto* and *de re*[3] -- and, ultimately, prove to be totally ineliminable. In fact, the problem is shown to remain even in indexicalized systems based on possible-worlds semantics (where propositions are functions from indices to truth-values). Here, of course, the indices are persons, times, and places rather than possible worlds, and beliefs are evaluated with reference to these explicitly indicated indices. Nonetheless -- and promising as this may sound -- Perry demonstrates that it makes no more progress towards eliminating indexicals than the other approaches examined.

Thus we are left with a need for agents that can make sense of indexicals in a meaningful way; and moreover, our system will need to actively use such terms as indexicals, rather than trying to replace them with fixed-meaning terms. In what follows, we outline a framework for doing so and describe some progress made.

We briefly note some other salient work in the problem of effectively drawing a self-other distinction. For example, Selmer Bringsjord (2015) devised a system that recognized its own utterances, but seemed to rely on external correlates of its speech rather than using knowledge about its initiation of speech. Hart & Scassellati (2011) used an approach similar to ours in some ways, but they employed motor-proprioceptive information rather than an "efference copy" of initiated action.

Why reasoning with *grounded* symbols is needed

We give a couple of example scenarios which are intended to emphasize the necessity of reasoning with grounded self-symbols even in relatively simple environments. The first of these relates to an issue we had in our lab when training one of our (Baxter) robots to respond to any command containing the word "Julia" by pointing to Julia (another robot) and saying "I see Julia and I'm pointing to her." We found that our robot was overly eager in performing this task; immediately after pointing at Julia she would point again and repeat "I see Julia and I'm pointing to her" -- and then continue to repeat the process until stopped. We eventually realized that our robot, on hearing herself speak, took this as a command to faithfully execute. This demonstrates that even a basic knowledge of actions must distinguish those initiated by others from self-initiated actions, even in the simplest contexts.

We can also imagine a pair of robots working on a task cooperatively. One of them, say "Alice", sees a robot arm moving rapidly towards her. She needs to very quickly understand whether it is her own arm moving under her own volition or another robot's arm that she is seeing – her response should very much depend on this determination. The extreme shortness of the time-scales involved also emphasizes that the self-other distinction needs to be drawn *pre-reflectively* – it cannot be the result of (relatively slow) conceptual determination but should rather be part of the very structure of the robot's cognition.

Ultimately, it will also be crucial for robots engaged in the real world to have a theory of mind. In the scenario above, if Alice determines that the robot arm she sees to be getting closer and closer belongs to another robot, she needs to make some determinations about what that other robot's intentions are so that she can respond appropriately. The most natural way to develop such a theory would seem to be for Alice to engage in introspection and reason about what kinds of motivations might cause her to behave in a certain way. Doing so will require Alice to have a well-developed subjective understanding of her own functioning.

Finally, we note that whatever system we use for dealing with indexicals (indeed *any* denoting terms) should be *grounded*. That is, the relation between the terms and what they refer to should be based on actual experience with those things. In particular, any use of the term "I" by our robots should be grounded in the robot's actual experiences -- rather than inferred through some correlate. One strong reason for this is that correlates tend to be imperfect and have a strong chance of being inadequate in a complex range of uncertain circumstances. As an example, if an agent recognizes its own speech acts on the basis of voice recognition software, then it is likely to have a great deal of difficulty recognizing that it is speaking if, say, for some physical reason (e.g. a hardware degradation) some characteristics of its voice change -- or if, perhaps, it hears a recording of its own voice; or again, if its fellow robots use the same speech synthesis software. Human agents, however, are quite robust in the face of all kinds of such variance.

Our Approach

Ultimately, we would like to have a full model of subjectivity that grounds our agent's use of "I" and "me" and "now." One fundamental component of this will be to develop agents with a grounded knowledge of their own actions (this is called *sense of agency* in the scientific literature; we will use the more deflationary term *representation of agency* to avoid the appearance of attributing phenomenal content to our agents). Our grounded representation of agency will lead easily to a grounded *representation of ownership*. That is, by understanding when it is acting, our agents can claim ownership of the sensory loci of its action (its body in movement and its voice in speech). While these two do not constitute a full grounded subjectivity, they are essential components of one (Tsakiris et al 2007).

[3]In which beliefs are seen as a relationship between a subject and an ordered pair of object plus property. See Perry's paper for a nice explication.

Our approach to representing agency and ownership is based on the neuroscience notion of *efference copy*. In humans, whenever a motor command is issued by the brain, a copy of that command is saved and fed into a forward model. This model makes a prediction about what sensory feedback should be expected from executing that command. This allows for two fundamental capabilities. In the first place, the expected sensory feedback can be subtracted from the incoming sensory input to the effect that self-generated stimuli are dampened – this is a prominent explanation for why people cannot tickle themselves (Blakemore et al, 2000). In the second place, a comparison between expected and received sensory feedback can allow an agent to do real time error correction. For example, if a motor command is issued to move 5 inches forward, and sensory feedback indicates that I have only moved forward 3 inches, then I can immediately apply extra force to correct for this.

The efference copy mechanism also provides our agents with a representation of agency: an agent can conclude that it has initiated actions in which its expectations match its received input. It also has the potential to provide for a representation of ownership by picking out the parts of the sensory input which correspond to the agent's actions (and are hence under its control). To do this effectively, a representation of the sensory input which allows for easily picking out which parts correspond to self-action is desirable. Ultimately we hope to train efficient hierarchical representations, although this is beyond the scope of the present work.

These low-level mechanisms for identifying self-actions and the bodily self are then taken to ground indexical symbols referring to an agent's actions and body. We employ an *active logic* reasoning agent (described below) to perform higher order reasoning over these representations. Our active logic engine also has built-in grounded temporal notions, so that all of the indexicals deemed fundamental by Perry will be accessible, including "I", and"now," along with derived indexicals such as "you" and "in 10 minutes."

Active Logic

At present we have a system that can generate an indexical symbol for its own speech based on an artificial efference copy. We are almost at the point of extending this work to handle basic motions in a grounded, self-aware way, again using an efference copy-like mechanism.

The current system provides for two-tiered symbols. For example, in the auditory domain there is a low-level module for producing speech and comparing auditory input to expected input in real-time. This module communicates with a higher level symbolic processing system and issues such tokens as "*doing(say(U),t)"* and "*speech_error(U, t)"* to indicate (respectively) that the agent is intentionally saying the utterance *U* at time *t* and that there was an error in doing so.

Our symbolic reasoning is done with ALMA, the *Active Logic MAchine*. Active logic is a formalism developed by our group specifically to deal with reasoning in dynamic and uncertain environments. In particular, it has following characteristics:

Diachronic: Active logic is time-sensitive in two senses. It is a temporal logic in that an active logic reasoner can reason *about* time. It has the further property that reasoning not only occurs *over* time but also is situated *in* time. That is, the reasoning process keeps track of when sentences enter in the knowledge base, how long it takes for a particular conclusion to be drawn, etc. In particular, active logic specifies a predicate *Now(i)* which is updated at each timestep and is true when *i* is the current timestep. *Now()* thus functions as a temporal indexical.

History-maintaining: Since sentences are effectively time-stamped, an active logic agent can keep track of what it used to believe and reason about that as well.

Paraconsistent: Since sentences represent beliefs at particular times (rather than eternal truths), contradictions are tolerated and can be handled. In particular, if the agent believes *S* and ~*S* at a particular time *t*, then it can later resolve that contradiction and believe, say, only *S* at time *t+k* for some *k>0*. (That is, if the agent comes to believe both a proposition is both true and false at the same time, it can resolve that contradiction and choose which to believe at a later time).

Metacognitive: All of the agent's reasoning -- including backward reasoning based on history -- could be done in one single stream of reasoning without relying on different reasoners at different levels.

With these features, active logics have been used for time-situated planning and execution (Purang, et. al., 1999); for reasoning about other agents (Kraus and Perlis, 1989); for reasoning about dialog (Perlis, et. al., 1998) including updating and using discourse context (Gurney et. al., 1998); and for implementing autonomous agency (Chong, et. al., 2002).

Progress to Date

In order to integrate ALMA with other software packages and robotic hardware, we use the Robot Operating System (ROS) for its message passing capabilities amongst other functions. Our system (running on a Baxter robot) currently has several interacting ROS nodes. The first of these is a wrapped version of ALMA which updates the ALMA knowledge base at regular intervals. This entails adding and removing sentences based on input to the node, and also deriving "single-step" consequences of the sentences that are already in the knowledge base (these correspond to single applications of modus ponens). By employing

standard ROS mechanisms, the other nodes in the system can both view the knowledge base in real-time and request that specific sentences be added or deleted.

A second node serves on behalf of the reasoning engine as a kind of effector. Essentially, if the ALMA node has concluded that it should perform an action (say *X*), then this effector node is be responsible for executing that action. It does this by scanning the knowledge base for sentences of the form "*action*(*X*)," and, when it finds one, it executes some code to undertake *X*. While this code is functioning, it replaces "*action*(*X*)" with "*doing*(*X, t*)," for those *t* during which the action is being undertaken and, finally, with "*done*(*X*)" when the code has completed successfully. In particular, this node handles actions of the form *say(U)* by passing the utterance *U* to the auditory sense of agency (ASOA) node, and replacing "*action*(*say*(*U*))" with "*doing*(*say*(*U*),*t*)" for those times t during which speech is occurring.

The ASOA node then represents speech acts with an artificial efference copy, comparing the heard speech with the expected speech and raising an error if there's a mismatch.

This allows our system to draw a meaningful self-other distinction for speech acts. For example, if a speaking action is currently labeled with "*doing(say(U))*," and the microphone is just picking up some similar-sounding speech, our reasoning engine can infer that it (itself) is the one causing the speech (and thus avoid reacting to it). Or on the other hand, if the mic does *not* pick up sounds similar to what the system takes itself to be doing, then this can sanction an inference to the effect that something is wrong, such as a damaged speaker or microphone. In particular, variants of the following active logic axioms are employed:

```
if(heardCommand(C)and not(doing(say(C)),
   processCommand(command(C)))

if(doing(say(U))and heardNothing,
   say("Is my microphone muted?"))

if(processCommand(pointJulia),
  action(pointToJulia) and
  action(say("I see Julia and am pointing at
              her.")
```

Here the idea is that if a command with text *C* is heard and it was not self-initiated, then we should process the corresponding command. In particular, if the command was to point to Julia, then we locate her, point to her and say that we are doing so. Also, if we are trying to speak and don't hear anything, we will ask the user if the microphone is muted.

Last year (Brody, Perlis and Shamwell 2015) we reported on a more primitive version of the above, in which the Baxter system was able to compare its outgoing wave-file with the incoming one, and thus refrain from reacting if the match was adequate. But this was an immediate reflexive act, not mediated by a reasoning process (ALMA).

In current work we are designing a similar facility – with another ROS node – to monitor and reason about self-initiated *movement*. Thus for instance, our Baxter can initiate locomotion (using a motorized base) or it can be moved by humans using a joystick. One immediate aim is to have the robot determine – via efference copy and sensory feedback – whether it is moving under its own control or not. This will however be more complicated than the ASOA case, for the sensory inputs (e.g., visual and proprioceptive) now will differ significantly in format from the efferent signal (motor commands).

References

Anderson, M. L., & Perlis, D. R. (2005). The roots of self-awareness. *Phenomenology and the Cognitive Sciences*, *4*(3), 297-333.

Blakemore, S. J., Wolpert, D., & Frith, C. (2000). Why can't you tickle yourself?. *Neuroreport*, *11*(11), R11-R16.

Bringsjord, S., et al. 2015. Real robots that pass human tests of self-consciousness. In *Proceedings of IEEE the 24th International Symposium on Robots and Human Interactive Communications.*

Brody, J., Perlis, D., & Shamwell, J. (2015, September). Who's Talking?—Efference Copy and a Robot's Sense of Agency. In *2015 AAAI Fall Symposium Series*.

Chong, W., O'Donovan-Anderson, M., Okamoto, Y., & Perlis, D. (2002, January). Seven days in the life of a robotic agent. In *Workshop on Radical Agent Concepts* (pp. 243-253). Springer Berlin Heidelberg.

Gurney, J., Purang, K., & Perlis, D. (1998). Updating discourse context with active logic.

Hart, J., and Scassellati, B. 2011. Robotic models of self. In: Cox and Raja, eds, *Metareasoning: Thinking About Thinking*. Cambridge, MA, MIT Press.

Kraus, S., & Perlis, D. (1989). Assessing others' knowledge and ignorance. In *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems* (pp. 220-225).

Lakoff, G. (1990). *Women, fire, and dangerous things: What categories reveal about the mind* (pp. 1987-1987). Chicago: University of Chicago press.

Perry, J. (1979). The problem of the essential indexical. *Noûs*, 3-21.

Perlis, D., Purang, K., Purushothaman, D., Andersen, C., & Traum, D. (1999). Modeling time and meta-reasoning in dialogue via active logic. In *Working notes of AAAI Fall Symposium on Psychological Models of Communication*.

Purang, K., Purushothaman, D., Traum, D., Andersen, C., & Perlis, D. (1999). Practical reasoning and plan execution with active logic. In *Proceedings of the IJCAI-99 Workshop on Practical Reasoning and Rationality* (pp. 30-38).

Tsakiris, M., Schütz-Bosbach, S., & Gallagher, S. (2007). On agency and body-ownership: phenomenological and neurocognitive reflections.*Consciousness and cognition*, *16*(3), 645-660.