RC24377 (W0710-074) October 15, 2007
Computer Science

IBM Research Report

Fusing Animals and Humans

Jonathan Connell
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Fusing Animals and Humans

Jonathan CONNELL
IBM T.J. Watson Research Center, Yorktown Heights NY, jconnell@us.ibm.com

Abstract. AI has many techniques and tools at its disposal, yet seems to be lacking some special "juice" needed to create a true being. We propose that the missing ingredients are a general theory of motivation and an operational understanding of natural language. The motivation part comes largely from our animal heritage: a real-world agent must continually respond to external events rather than depend on perfect modeling and planning. The language part, on the other hand, is what makes us human: competent participation in a social group requires one-shot learning and the ability to reason about objects and activities that are not present or on-going. In this paper we propose an architecture for self-motivation, and suggest how a language interpreter can be built on top of such a substrate. With the addition of a method for recording and internalizing dialog, we sketch how this can then be used to impart essential cultural knowledge and behaviors.

Keywords. Agent architecture, Motivation, Learning, Language, Clicker training

1. What is intelligence?

In order to achieve AI it helps to know how the end product will be evaluated. In particular, we focus on "perceived intelligence" – not what might count as the true Platonic ideal of intelligence, but what properties an external observer might take as evidence that there is something there. In the natural world the layered ordering shown below is what is typically observed. There are arguably more organisms that are aware than ones with personality (e.g. ants), and there are animals that are social without having much in the way of abstraction capabilities (e.g. guinea pigs). Yet it is not clear that the lower levels are an absolute prerequisite for the higher levels. For instance, there are many computer programs that deal in abstractions but have no personality.

Criteria for *Perceived* Intelligence

1. **Animate** – Coordinated movement, many degrees of freedom
2. **Aware** – Responds and changes actions based on environment change
3. **Personality** – Individuals have different likes / dislikes, preferences learned
4. **Social** – Aware of social order, use other beings as agents
5. **Abstract** – Conceptualize situations remote in space and time, planning
6. **Communicative** – Express internal ideas and ingest situational descriptions

To better understand the above criteria imagine applying them to a robot toy. Obviously if it just sits there and does nothing it is pretty boring – it has failed criterion 1. Now suppose that it can zoom around at high speed but constantly runs smack into things. It seems vaguely alive but not very smart at all – it has failed criterion 2. The next step is to exhibit some sort of personal preferences for things, activities, situations, or people. The Roomba vacuum cleaner, created as a tool, fails criterion 3. By contrast, Furby, an animatronic toy that complains about being turned upside down, succeeds. Furby also partially passes criterion 4 because it is forever badgering its owners to "Feed me! Yummm." and having them comply. Although annoying, Furby is arguably closer to what we want in a true AI than many other artifacts (or at least it has what most other AI programs lack).

The next two criteria seem more applicable to animals than robots. There is a lot of literature on parrots, crows, pigs, dogs, dolphins, monkeys, and chimps concerning tool use, sequential tasks, and delayed reward scenarios – all activities requiring some proficiency at criterion 5. By contrast, criterion 6 seems largely limited to humans. Much of this is tied to language, something animals have been able to acquire only to a limited extent [1]. In many ways language is absolutely required for being human. People do not exist as singletons: we are all part of a progressively more tightly coupled "super-organism". Language is the basis for this coupling – it allows us to have knowledge of things our bodies have never experienced directly, particularly things remote in space or time (cf. criterion 5). To be human is to be able to participate in this cultural super-organism, and to do that requires language.

Looking to existence proofs for inspiration, when watching an animal that has undergone clicker training it really feels like "someone" is there. The whole training paradigm provides a nice mechanism for autonomous goals and learning [2]. On the other hand, there are programs like BORIS [3] which interprets narratives and handles natural language questions. This exhibition of deep language understanding is hard to ignore. Moreover, BORIS allows its deductions to be tailored by instruction, something that feels very human. To get the best of both we propose fusing animals and humans.

2. The animal part – general purpose motivation

The animal part is largely about motivation: why animals do what they do and when they do it. Yet it is important to realize that animals spontaneously do things all the time. Pets do not await a command from their owners, and wild animals obviously have no external agency calling the shots. In this way they are very different from contemporary computer programs which amplify and elaborate on imperatives supplied by their users. Moreover, although goals can be internally generated rather than supplied from the outside, it seems unlikely that animals are always operating in a goal-driven manner, except in the loosest sense. That is, much of what animals seem to do is routine or reflexive – there is no specific articulated goal which they are pursuing. Ongoing activities such as foraging, grooming, and migration can not be traced back through a logical chain of reasoning to a concrete goal proposition. To mirror this observation, the motivation system proposed here has only a loose coupling between goals and actions. As shown in Figure 1, the bulk of the activity is controlled reactively using situation-action policies [4, 5]. Some of these policies are active all the time, while others can be switched in or out depending on the situation [6]. Much of the

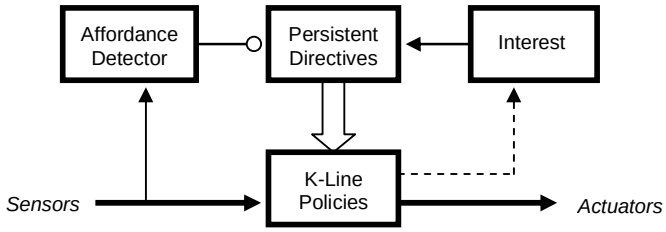

Figure 1. A loosely coupled architecture for self-motivation.

switching of polices is governed by a set of persistent directives. In this respect the action component of the architecture is similar to Minsky's K-lines [7].

Yet where do these policies and directives come from? One possible answer is that we can exploit <S, E, A> triples. Here E is an exciting event, S represents the observed situational context, and A captures the current actuator settings. These <S, E, A> triples are not recorded all the time, but only when something "interesting". The rules governing their formation and use are detailed below:

S & E & A & I(E) → <S, E, A> <S, E, A> & S & I(E) → D(E)

D(E) & <S, E, A> & S → A D(E) & <S, E, A> → I(S)

There are three uses for <S, E, A> triples. First they can function as affordance detectors – determining when the environment may be offering a useful opportunity to the agent. In this case the recorded S is used to predict the interesting additional observation E that might already hold, or may be forced to become present through some action. Although E was originally "interesting" when it was recorded, the organism's desires and needs routinely change over time. Thus each such proposed affordance is evaluated against the current "interest" metric (I). Then, if sufficiently stimulating, a desire (D) for this event is latched in as one of the persistent directives that control the behavioral policies. The other major use for <S, E, A> triples is as part of a policy. If the E part matches one of the current directives then when situation S occurs the agent will be prompted to try action A. Triples can also be used for loose backwards chaining. If E is one of the active directives then the agent should become interested in situation S. In this way it can serendipitously learn how to accomplish subgoals in the course of an activity (whether or not the associated action is taken).

As an example, suppose you are walking along the shore of a pond and there is a sudden splash as a frog jumps into the water. Since sudden noises are intrinsically interesting this would prompt the formation of a triple like <pond, splash, walk-along>. Now every time you see a pond the splash will be brought to mind. If this is still an interesting occurrence you will latch it in as something you desire to happen. This in turn activates one or more policies that bias you into doing things that might lead to a splash, such as if you see a pond you should walk along its shore. The desire for splashing may also activate other triples such as <pond & rock, splash, drop-rock>. Since the situation part of this is not yet fulfilled its action is not performed. Yet the set of interesting things will be expanded temporarily to include rocks, prompting background learning about where to find rocks (for this or any other purpose).

3. The human part – a language interpreter

The bulk of human-level intelligence is due to acquired cultural knowledge, at least observationally. If a person suffers amnesia and forgets his family and culture, in many ways he has been erased as the person he previously was. Thus it is essential for an AI to have the ability to learn patterns of cultural interaction. Language is both the key for learning this and the medium by which proficiency in culture is evidenced [8]. It is a sort of mind-sharing "telepathy" that makes an agent part of a cultural super-organism and allows its peers to "program" it to act in certain ways in specific situations [13].

At the bottom, language can be viewed as a general-purpose scripting system that invokes sensory/motor specific subroutines as needed. One way to become proficient in a language is to build an interpreter that allows the agent to receive descriptions and instructions via language then evaluate these structures to operationalize the knowledge. The starting point for this is to ground the fragmentary acoustic utterances in sensory and actuator experiences [9, 10]. As a prerequisite the system must have some innate mechanisms of parsing the world into physical objects and temporal events. It also needs some reflexes which will exercise all the actuator options and an initial good/bad reinforcement signal as feedback. An intrinsic means to guide attention is also very useful. All these abilities can be bootstrapped to higher levels of sophistication with a method like clicker training (i.e. operant conditioning with staging and shaping) [11].

Once a basic level of linguistic grounding has been achieved, learning new concepts and tasks becomes much easier [12]. For instance, one can say "No, this is a moth not a butterfly. Look at its fuzzy antennae." This explanation is much simpler and faster than showing the agent tens of contrasting examples and having it guess which features are relevant to the classification. Similarly, language can be used to directly impart routines for accomplishing sequential tasks, such as opening a jar of pickles. Instead of letting the agent fiddle with it for hours and shouting "Hurray!" when the top finally pops off, language allows the teacher to say "Hold the jar in your left had, grasp the top with your right hand, and twist hard". If this sequence of verbal directions is memorized and played back through the interpreter, the agent has essentially learned a plan for how to do the task in one shot (perhaps without ever actually performing it!).

Recording is one aspect of internalizing dialog [8], but there is more. The agent may start out with a running stream of remembered commentary: "Okay, the stop sign is coming up. Slow down and watch for other cars at the intersection. It the car on the cross street arrives first he gets to go first. It looks like there is no one around, so you can go now …" Yet this scaffolding can eventually be compiled out [14] until all that is left are the relevant behavioral policies. In particular, *compiling out* responses can yield internal deduction with out requiring any special-purpose "declarative assertion" action module. The agent predicts what it is going to say and what this will sound like. It can then get the effect of the interpreter running on that acoustic input directly without actually speaking out loud. That is, the middle two steps below get compiled out.

(see: shaggy animal → say: "It's a dog") → (hear: "It's a dog." → represent: dog)

see: shaggy animal → represent: dog

Similarly, *compiling in* questions can yield behaviors for directing attention and more thoroughly processing input. Here a teacher-narrated sequence about identifying birds is condensed by means of removing the dialog components in the middle.

(see: bird → hear: "It's a bird!") → (hear: "What shape is its beak?" → look: at beak)

see: bird → look: at beak

4. Discussion

Our basic recipe for intelligence has three steps. First, the animal part provides a substrate for learning an interpreter for language. Second, language in turn unlocks a method for rapid transmission of concepts and behavior patterns. Third, competent performance within such a defined culture comprises the human part. So why not just build a language interpreter and jettison the animal heritage? Presumably the agent could be trained to be properly curious, ask questions when appropriate, and initiate further information gathering activities when needed. Perhaps the answer is that the animal part lets the agent fall back on *weak methods* [15] when its programming fails. Instead of just crashing and staring blankly ahead until it receives new top level instructions, if the agent maintains the proper set of directives and interest biases it can muddle through to some state where the more detailed program can pick up once again.

In this paper we proposed six criteria for perceived intelligence, sketched an architecture for the lower three, and argued that top one was largely a function of language. What then of the middle two? It has been posited that deep social intelligence is another hallmark of humanity [16]. Yet some interesting work [17] has shown how a beginning theory-of-mind can be built on top of mechanisms like shared gaze, all implemented in an animal-like architecture. Abstract thought, the remaining criterion, may need some pre-existing mechanism (e.g. spatial navigation) for language to latch on to. Then again, it may be that abstract abilities like planning can emerge as a natural outgrowth of remembered or internalized activity patterns (cf. opening the pickle jar).

References

[1] E. Kako, "Elements of syntax in the systems of three language-trained animals", *Animal Learning & Behavior*, 27(1), pp. 1-14, 1999.
[2] L. Saksida, S. Raymond, and D. Touretsky, "Shaping Robot Behavior Using Principles from Instrumental Conditioning", *Robotics and Autonomous Systems*, 22 (3/4):231, 1998.
[3] Michael Dyer, In-Depth Understanding, MIT Press, 1983.
[4] R. Brooks, "A Robust Layered Control System for a Mobile Robot", *Journal of Robotics and Automation*, RA-2, pp. 14-23, 1986.
[5] J. Connell, Minimalist Mobile Robotics, Academic Press, 1990.
[6] J. Connell and P. Viola, "Cooperative Control of a Semi-Autonomous Mobile Robot", *Proc. of the IEEE Conf. on Robotics and Automation* (ICRA-90), pp. 1118-1121, 1990.
[7] M. Minksy, "K-Lines: A Theory of Memory", MIT AI Memo 516, 1979.
[8] Lev Vygotsky, Thought and Language, MIT Press, 1962.
[9] L. Steels and F. Kaplan, "AIBO's first words", *Evolution of Communication*, 4(1), pp. 3-32, 2001.
[10] D. Roy, "Semiotic Schemas", *Artificial Intelligence*, 167(1-2), pp.170-205, 2005.
[11] Karen Pryor, Don't Shoot the Dog, Simon & Schuster, 1984.
[12] J. Connell, "Beer on the Brain", *Proc. of the 2000 AAAI Spring Symposium, My Dinner with R2D2: Natural Dialogues with Practical Robotics Devices*, pp. 25-26, 2000.
[13] B. F. Skinner, Verbal Behavior, Appleton-Century Crofts, 1957.
[14] John R. Anderson, Rules of the Mind (Chapter 4), Lawrence Erlbaum, 1993.
[15] J. Laird, A. Newell, and P. Rosenbloom, "Soar: An Architecture for General Intelligence", *Artificial Intelligence*, 33(1), pp. 1-64, 1987.
[16] E. Herrmann, J. Call, M. Hernández-Lloreda, B. Hare, and M. Tomasello, "Humans Have Evolved Specialized Skills of Social Cognition", *Science*, vol. 317, pp. 1360–1366, 2007.
[17] B. Scassellati, "Theory of Mind for a Humanoid Robot", *Autonomous Robots*, vol. 12, pp. 13–24, 2002.