The Emergence of Representation in Autonomous Embodied Agents

Mark H. Bickhard [1]
Cognitive Science
17 Memorial Drive East
Lehigh University
Bethlehem, PA 18015
610-758-3633
mhb0@lehigh.edu
http://www.lehigh.edu/~mhb0/mhb0.html

A problem of action selection emerges in complex — and even not so complex — interactive agents: What to do next? The problem of action selection occurs equally for natural and for artificial agents — for *any* embodied agent. The obvious solution to this problem constitutes a form of representation — interactive representation — that is arguably *the* fundamental form of representation. More carefully, interactive representation satisfies a criterion for representation that no other model of representation in the literature can satisfy, or even attempts to address: the possibility of system detectable representational error. It also resolves and avoids myriad other problematics of representation, and integrates or opens the door to many additional mental processes and phenomena, such as motivation.

Autonomous Agents and Interaction Selection

In sufficiently simple interactive systems, there may be only one thing to do, and it is simply done all of the time. Slightly more complex would be the case in which multiple actions or interactions are possible, but only one of them is triggered in any possible combination of system state and input. With only minimal interactive complexity, however, there will be conditions in which multiple interactions are possible, and a simple triggering of one of those possibilities will be insufficient. This will occur, for example, if the system-environment circumstances have not occurred before, or have occurred too few times to have selected for a simple triggering relation, or when the environment is insufficiently predictable for it to be determinate what the system outcomes of the interactions will be (Bickhard & Terveen, 1995). A given frog tongue flick may — or may not — catch a fly.

Action selection in such cases poses its own special problems. In particular, whatever presuppositions are involved in the general strategy of interaction selection may, in particular cases, turn out to be wrong. In general, frogs flick tongues in order to catch flies (or worms, ...), but such outcomes are not guaranteed. In some cases, the advisable strategy would be to simply try again when the interaction doesn't work, while in other cases the advisable strategy might be to switch to a different next interaction selection.

But implementing *any* such design constraints requires that the system be able to detect when the "desired" outcomes have occurred, or have not occurred. That, in turn, requires that those "desired" outcomes be somehow functionally indicatable, and detectable. In other words, the agent has to be able to differentiate between interaction outcome success and interaction outcome failure, and this requires that it be able to be functionally sensitive to information about interaction outcomes.

There are many traps in the way to an implementation of these notions. For example, if the "desired" outcomes are certain environmental conditions — perhaps a new position relative to some food source — then those environmental conditions will have to be *represented* by the system, and failures to achieve those environmental conditions will have to be detectable. But representation is what we are aiming to model in the first place, so this approach initiates a fatal circularity.

I will not dwell on the multiple dangers involved in this modeling

(Bickhard, 1993; Bickhard & Terveen, 1995), but will move more directly to the solution. The relevant interaction outcomes must be *internal* so that they do *not* have to represented, but can be directly functionally accessible to the system. Such internal interaction outcomes must be *anticipated* (Bickhard, 1993; Brooks, 1994) in such a way that success and failure of those anticipations can be detected.

There are multiple architectures within which the functions of anticipation and success/failure detection can be implemented (Bickhard & Terveen, 1995). For my current purposes, I will focus only on simple and familiar architectural principles of states and pointers. In particular, interaction outcomes will be system *states*, so anticipations of system states is what is required. Anticipations, in turn, require only *indications* of such system states associated with the relevant *interactions* that might or should yield those states. And, finally, such indications can be accomplished with simple *pointers*. Interaction possibilities can be indicated with pointers to the subsystems or system conditions that would engage in those interactions if initiated, while associated outcomes can be indicated with associated pointers to the relevant internal states. So long as the system can detect whether or not those indicated states have been achieved, all of the functional requirements for this form of interaction selection have been met.

Selecting interactions on the basis of the indicated outcomes of those interactions permits many powerful phenomena. It permits strategic selection of interactions *toward* particular outcomes — *goal* directedness. It permits *error guided strategies* of interaction. It permits error guidance of representational *learning*, so long as the system is capable of learning at all. Any complex autonomous agent will require such forms of interaction selection. Note, further, that interaction selection is the fundamental problem of motivation (Bickhard forthcoming-b; Bickhard & Terveen, 1995; Cherian & Troxell, 1995a; Maes, 1990a, 1990b, 1991, 1992, 1994)

Representation

Autonomous agents encounter problems of interaction selection. A natural solution to those problems involves internal indications of internal interaction outcomes. The central claim of this paper is that such indications constitute a form of representation, arguably the fundamental form.

I will support this claim in three ways. The first two supports address two properties of representations that this model accounts for: 1) content and 2) truth value of that content. The third points out that the interactive model provides an account for a fundamental criterion for original representation that is not and cannot be addressed by any standard model in the contemporary literature: the possibility of *system detectable* representational error.

Representational Content. The claim that indications of interaction outcomes can constitute representation poses a number of questions. A central question is: What constitutes representational content in this model? The answer has, in fact, already been adumbrated above: an indication of an interaction and its associated outcomes makes presuppositions about the environment. In particular, those presuppositions will hold in some circumstances and not hold in others, and, correspondingly, the indicated outcomes will occur in some — the presupposed — circumstances and not occur in others. Such indications involve implicit presuppositions of the "appropriate" properties of the environment for those indicated outcomes to occur.

These presuppositions provide a model for representational content. The presupposed interactive properties of the environment are what are represented of that environment. They are what are predicated of that environment: "This environment has interactive properties appropriate to these indicated interaction outcomes." The content in this model is *implicit* rather than explicit (Bickhard, 1993; Bickhard & Terveen, 1995; Brooks, 1991b), presupposed rather than portrayed. This is a major strength of the model in that it avoids many aporetic problems concerning representation (Bickhard, 1993; Bickhard & Terveen, 1995). It also poses its own problems — for example, how then do we account for clear cases of explicit content? How can representations of objects be modeled, not just representations of interactive properties? How about representations of abstractions, such as numbers? Such questions lead into further elaborations and explorations of the interactive model. I have addressed many of these elsewhere (Bickhard, 1980, 1993, forthcoming-a, forthcoming-b; Bickhard & Richie, 1983; Bickhard & Terveen, 1995 and other references therein). At this point, I simply want to point out that the interactive model does provide an account of content. The epistemological world is constructed out of such content (Bickhard, 1993; Bickhard & Terveen, 1995; Hooker & Christensen, in preparation; Kuipers, 1988; Kuipers & Byun, 1991; Mataric, 1991; Nehmzow & Smithers, 1991, 1992).

Truth Value. Furthermore, those presuppositions — those implicit contents — can be *false*, in which case the outcomes will not occur. Or, they can be true, in which case the outcomes will occur. That is, the actual environment is the truthmaker for the implicit content — and, consequently, the interactive model of representational content can account for truth value, one of the fundamental properties of representation.

System Detectable Representational Error

Still further, interactive representational truth value can detected, at least fallibly, by the interactive system itself. The outcome indications can be falsified by and for the system, when it detects that the indicated outcome did not occur. That constitutes an falsification of the presupposed properties, the implicit content, about the environment in which the interaction occurred. It is a falsification that is accomplished by the system itself.

Note that, unless such falsifications can occur by and for the agent itself, error guidance is not possible, whether interactive error guidance or learning error guidance. Error guidance, in turn, is required in any complex interactive system because the certitude required for, say, simple triggering, is not possible.

System detectable error, however, is not possible for standard models of representation in the literature. Standard models have difficulty accounting for any kind of error, and do not even address system detectable error. Error detection for Fodor (1987, 1990), Dretske (1981, 1988), and Millikan (1984, 1993), for example, requires comparisons between what actually triggers an input state and what "ought" to trigger that state. What "ought" to have triggered such a state, in turn, is determined by considerations of evolutionary histories and counterfactual asymmetric dependencies between classes of possible such triggers. The necessary comparisons, then, are possible, if at all, only for an observer of the system, external to the system, that might be in a position to determine the relevant histories and dependencies. In particular, they cannot be determined by and for the system itself. The frog knows nothing about the evolutionary history of its fly detectors, nor about any dependencies, asymmetric or otherwise, between fly detections and BB detections. Standard models cannot account for system detectable representational error; standard models cannot account for error guided interaction or learning; standard models cannot account for representation for the agent itself. The interactive model can. (Furthermore, interactive representational content can "ground" internal symbols as stand-ins for those contents — Bickhard & Terveen, 1995.)

Note that these standard models all look externally and backward in time, while the interactive model looks internally and forward in time (Bickhard & Terveen, 1995). It is such backward-looking "spectator" models that arise naturally in symbol manipulation, information theoretic, or connectionist models, but such an approach has never been able to solve the basic problems of representation. It has, rightly, yielded a reaction against the reality or usefulness of the very notion of representation (Beer, 1990, 1995; Brooks, 1991a, 1991c; Prem, 1995). Interactive representation constitutes an alternative.

Representation and Design

Interactive representation already exists. It can be found in relatively simple organisms (not to mention complex organisms) and in some contemporary robots (Brooks, 1994; Cherian & Troxell, 1995a, 1995b). It solves a basic design problem, and provides guidance for more sophisticated versions of such problems and their solutions; it solves and dissolves many philosophical problems (Bickhard, 1993; Bickhard & Terveen, 1995). Conversely, the study of robots and autonomous embodied agents *inherently* encounters the fundamental problems of representation and epistemology, while passive system approaches, whether computational or connectionist, do not (Bickhard, 1982; Bickhard & Terveen, 1995). Interactive dynamic systems — embodied agents — are the correct locus for the study and design of mentality (Beer, 1995; Bickhard & Terveen, 1995; Hooker, Penfold, Evans, 1992; Malcolm, Smithers, Hallam, 1989; Port & van Gelder, 1995; Steels, 1994).

It also, however, initiates a progression through many additional mental properties, and thereby guides further understanding of those properties both in natural systems and in the design of artificial systems. These include, for example, motivation, perception, learning, emotions, consciousness and reflexive consciousness, language, rationality, and so on (Bickhard, 1980, 1993, forthcoming-a, forthcoming-b; Bickhard & Richie, 1983; Bickhard & Terveen, 1995; Hooker, 1995; Kinsbourne, 1988). Interactive representation constitutes the interface of the emergence of mental properties in functional interactive system properties.

References

Beer, R. D. (1990). *Intelligence as Adaptive Behavior*. Academic.

Beer, R. D. (1995). Computational and Dynamical Languages for Autonomous Agents. In R. Port, T. J. van Gelder (Eds.) *Mind as Motion: Dynamics, Behavior, and Cognition*. MIT.

Bickhard, M. H. (1980). *Cognition, Convention, and Communication*. New York: Praeger.

Bickhard, M. H. (1982). Automata Theory, Artificial Intelligence, and Genetic Epistemology. *Revue Internationale de Philosophie, 36*(142-143), 549-566.

Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, *5*, 285-333.

Bickhard, M. H. (forthcoming-a). Critical Principles: On the Negative Side of Rationality. In Herfel, W., Hooker, C. A. (Eds.) *Beyond Ruling Reason: Non-formal Approaches to Rationality.*

Bickhard, M. H. (forthcoming-b). Is Cognition an Autonomous Subsystem? In O'Nuallain, S. (Ed.) *Computation, Cognition, and Consciousness*. John Benjamins.

Bickhard, M. H., Richie, D. M. (1983). *On the Nature of Representation: A Case Study of James J. Gibson's Theory of Perception*. New York: Praeger.

Bickhard, M. H., Terveen, L. (1995). *Foundational Issues in Artificial Intelligence and Cognitive Science — Impasse and Solution*. Amsterdam: Elsevier Scientific.

Brooks, R. A. (1991a). Intelligence without Representation. *Artificial Intelligence, 47*(1-3), 139-159.

Brooks, R. A. (1991b). Challenges for Complete Creature Architectures. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (434-443). MIT.

Brooks, R. A. (1991c). New Approaches to Robotics. *Science, 253*(5025), 1227-1232.

Brooks, R. A. (1994). Session on Building Cognition. Conference on The Role of Dynamics and Representation in Adaptive Behaviour and Cognition. University of the Basque Country, San Sebastian, Spain, December 9, 1994.

Cherian, S., Troxell, W. O. (1995a). Intelligent behavior in machines emerging from a collection of interactive control structures. *Computational Intelligence*, *11*(4), 565-592. Blackwell Publishers. Cambridge, Mass. and Oxford, UK.

Cherian, S., Troxell, W. O. (1995b). Interactivism: A Functional Model of Representation for Behavior-Based Systems. In Morán, F., Moreno, A., Merelo, J. J., Chacón, P. *Advances in Artificial Life*: Proceedings of the Third European Conference on Artificial Life, Granada, Spain. (691-703). Berlin: Springer.

Dretske, F. I. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT.

Dretske, F. I. (1988). *Explaining Behavior*. Cambridge, MA: MIT.

Fodor, J. A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A Theory of Content*. Cambridge, MA: MIT Press.

Hooker, C. A. (1995). *Reason, Regulation, and Realism: Towards a Regulatory Systems Theory of Reason and Evolutionary Epistemology*. SUNY.

Hooker, C. A., Penfold, H. B., Evans, R. J. (1992). Towards a Theory of Cognition Under a New Control Paradigm. *Topoi, 11*, 71-88.

Hooker, C. A., Christensen, W. (in preparation). Very Simple Minds.

Kinsbourne, M. (1988). Integrated Field Theory of Consciousness. In A. J. Marcel, E. Bisiach (Eds.) *Consciousness in Contemporary Science*. (239-256). Oxford University Press.

Kuipers, B. J. (1988). The TOUR Model: A Theoretical Definition. From Kuipers, B. J., Levitt, T. Navigation and Mapping in Large-Scale Space. *AI Magazine, 9(*2), 25-43.

Kuipers, B. J., Byun, Y. (1991). A Robot Exploration and Mapping Strategy Based on a Semantic Hierarchy of Spatial Representations. *Robotics and Autonomous Systems, 9*, 47-63.

Maes, P. (1990a). *Designing Autonomous Agents*. MIT.

Maes, P. (1990b). Situated Agents Can Have Goals. In P. Maes (Ed.) *Designing Autonomous Agents*. (49-70). MIT.

Maes, P. (1991). A Bottom-Up Mechanism for Behavior Selection in an Artificial Creature. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (238-246). MIT.

Maes, P. (1992). Learning Behavior Networks from Experience. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (48-57). MIT.

Maes, P. (1994). Modeling Adaptive Autonomous Agents. *Artificial Life, 1*, 135-162.

Malcolm, C. A., Smithers, T., Hallam, J. (1989). An Emerging Paradigm in Robot Architecture. In T Kanade, F.C.A. Groen, & L.O. Hertzberger (Eds.) *Proceedings of the Second Intelligent Autonomous Systems Conference*. (pp. 284-293). Amsterdam, 11--14 December 1989. Published by Stichting International Congress of Intelligent Autonomous Systems.

Mataric, M. J. (1991). Navigating With a Rat Brain: A Neurobiologically-Inspired Model for Robot Spatial Representation. In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (169-175). MIT.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT.

Millikan, R. G. (1993). *White Queen Psychology and Other Essays for Alice*. MIT.

Nehmzow, U., Smithers, T. (1991). Mapbuilding Using Self-Organizing Networks in “Really Useful Robots.” In J.-A. Meyer, S. W. Wilson (Eds.) *From Animals to Animats*. (152-159). MIT.

Nehmzow, U., Smithers, T. (1992). Using Motor Actions for Location Recognition. In F. J. Varela, P. Bourgine (Eds.) *Toward A Practice of Autonomous Systems*. (96-104). MIT.

Port, R., van Gelder, T. J. (1995) *Mind as Motion: Dynamics, Behavior, and Cognition*. MIT.

Prem, E. (1995). Grounding and the Entailment Structure in Robots and Artificial Life. In Morán, F., Moreno, A., Merelo, J. J., Chacón, P. *Advances in Artificial Life*: Proceedings of the Third European Conference on Artificial Life, Granada, Spain. (39-51). Berlin: Springer.

Steels, L. (1994). The Artificial Life Roots of Artificial Intelligence. *Artificial Life, 1*(1), 75-110.

[1] Thanks are due to the Henry R. Luce Foundation for support during the preparation of this paper.