Emergent Intentionality in Perception-Action Subsumption Hierarchies

David Windridge

Perception-Action learning is a novel paradigm in robotics that aims to address significant deficits in traditional approaches to embodied computer vision [1]. In particular, in the conventional approach to autonomous robotics, a computer vision system will typically be employed to build a model of the agent's environment *prior* to the act of planning the agent's actions within the domain. Visual data arising from these actions will then typically be used to further constrain the environment model, either actively or passively (in active learning the agent actions are driven by the imperative of reducing ambiguity in the environment model).

However, it is apparent that there exists in this approach, a very wide disparity between the visual parameterization of the agent's domain and its action capabilities within it [2]. For instance, the parametric freedom of a front-mounted camera will typically encompass the full intensity ranges of the RGB channels of each individual pixel of the camera CCD, such the the range of *possible* images that might be generated in each time-frame is of an extremely large order of magnitude (of course, only a minuscule fraction of this representational space would ever be experienced by the agent). On the other hand, the agent's motor capability is likely to be very much more constrained (perhaps consisting of the possible Euler angle settings of the various actuator motors). This disparity leads directly to the classical problems of *framing* [3] and *symbol grounding* [4] (note that this observation is not limited purely to vision based approaches - alternative modalities such as LIDAR and SONAR would also exhibit the same issues).

Perception-Action (P-A) learning aims to overcome these issues, adopting as its motto, 'action *precedes* perception' [5, 6]. By this it is meant that, in a strictly formalisable sense, actions are conceptually prior to perceptions; i.e. perceptual capabilities should depend on action capabilities and not vice versa.

Importantly, this permits *Cognitive Bootstrapping* [7], the bootstrapping of an autonomous agent's representation framework at the same time as its representation of the world in terms of that framework. To do this, the Perception-Action learning agent proceeds by randomly sampling its motor space ('motor babbling'). For each motor action that produces a discernible perceptual output in the initial sensor space S (typically a RGB camera), a percept $p_i \in S$ is greedily allocated. The agent thus progressively arrives at a set of novel percepts that relate directly to the agent's action capabilities in relation to the constraints of the environment (i.e. the environment's *affordances* with respect to the embodied agent). This set of experimental data points $\cup_i p_i \subset S$ can, in principle, be generalized-over so as to create a *manifold* of high-level action capabilities; i.e a set of affordance *hypothesizes* - the aspects of the agent's world that the agent is plausibly able to change. Critically, because these possibilities represent a constrained subset, they are susceptible to parametric compression. Furthermore, this parametric compression in the action domain (corresponding to the bootstrapping of a higher level action) necessarily corresponds to a parametric compression in the perceptual domain (P-A learning enforces a bijective relation $\{actions\} \rightarrow \{percept_{\text{initial}}\} \times \{percept_{\text{final}}\}$ such that each hypothesizable action has a unique, discriminable outcome [7, 8, 9]). A higher level percept domain (e.g. 'objects') is thus created along with

each higher level action (e.g. 'translate object').

Very often compressibility will be predicated on the discovery of *invariances* in the existing perceptual space with respect to randomized exploratory actions. Thus, for example, an agent might progress from a pixel-based representation of the world to an object-based representation of the world via the discovery that certain patches of pixels *retain their (relative) identity* under translation, i.e. such that it becomes far more efficient to represent the world in terms of indexed objects rather than pixel intensities (though the latter would, of course, still constitute the base of the representational hierarchy). This particular representational enhancement can represent an enormous compression [10]; a pixel-based representation has a parametric magnitude of P^n (with P and n being the intensity resolution and number of pixels, respectively), while an object-based representation typically has a parametric magnitude of $\sim n^o$, $o << n$, where o is the number of objects.

When such a high level perceptual manifold is created it permits proactive sampling - the agent can propose actions with perceptual outcomes that have not yet been experienced by the agent, but which are consistent with its current representational model (this guarantees falsifiability of both the perceptual model as well as the generalized affordance model). Perception-Action learning thus constitutes a form of active learning: randomized selection of perceptual goals within the hypothesized perception-action manifold leads more rapidly to the capture of data that might falsify the current hypothesis than would otherwise be the case (i.e. if the agent were performing randomly-selected actions within in the original motor domain). Thus, while the system is always 'motor babbling' in a manner analogous to the learning process of infant humans, the fact of carrying out this motor babbling in a higher-level P-A manifold means that the learning system as a whole more rapidly converges on the correct model of the world.

This P-A motor-babbling activity can take place in *any* P-A manifold, of whatever level of abstraction; we may thus, by combining the idea of P-A learning with Brooke's notion of task subsumption, conceive of a *hierarchical Perception-Action learner* ([11]), in which a vertical representation hierarchy is progressively constructed for which randomized exploratory motor activity at the highest level of the corresponding motor hierarchy would rapidly converge on an ideal representation of the agent's world in terms of its affordance potentialities. Such a system would thus converge upon both a model of the world, and an ideal strategy for representation of that world in terms of the learning agent's action capabilities within it.

Perceptual goals thus exist at all levels of the hierarchy, and the subsumptive nature of the hierarchy means that goals and sub-goals are scheduled with increasingly specific content as the high-level abstract goal is progressively grounded through the hierarchy. (Thus, as humans, we may conceive the high-level intention 'drive to work', which in order to be enacted, involves the execution of a large range of sub-goals with correspondingly lower-level perceptual goals e.g. 'stay in the center of the lane', etc).

Within the Perception-Action learning agent, the environment has thus "become it *own* representation", [12], representing a significant compression of the information that an agent needs to retain. This relates to the issue of symbol grounding, a seminal problem in the conceptual underpinning of the classical approach to machine learning [4]. The problem arises when one attempts to relate an abstract symbol manipulation system (it was a common historical assumption that computational reasoning would center on first-order logic deduction) with the stochastic, shifting reality of sensor data. In hierarchical P-A learning the problem is eliminated by virtue of the fact that symbolic representations are *abstracted from the bottom-up* [13, 14, 15, 5]. They are thus always intrinsically grounded (for an example of utilization of first-order logic induction within a subsumption hierarchy see [7]).

The subsumption hierarchy is thus typically characterized by continuous stochastic relationships on the lower levels and discrete symbolic manipulation at the higher levels. In such a cognitive bootstrapping system, motor-babbling at the top of the representation hierarchy involves the spontaneous scheduling of perceptual goals and sub-goals at the lower level of the hierarchy in a way that (as the hierarchy becomes progressively deeper) looks increasingly *intentional* (a

phenomenon that is readily apparent in the development of motor movement of human infants as schema abstraction takes place [16]). For instance, the intuition of a generalized percept category *container* correlates with the attempt to falsify this notion via the repeated placing of a variety of objects into a variety of containers). Such high-level percept-motor babbling can, in principle, be detected via an appropriate classification system.

Moreover, this percept-action relationship may be modeled in reverse to characterize human intentional behavior; consider how, as humans we typically represent our environment when driving a vehicle. At one level, we internally represent the immediate environment in metric-related terms (i.e. we are concerned with our proximity to other road users, to the curb and so on). At a higher level, however, we are concerned primarily with *navigation*-related entities (i.e how individual roads are *connected*). That the latter constitutes a higher hierarchical level, both mathematically and experientially, is guaranteed by the fact that the topological representation *subsumes*, or supervenes upon, the metric representation; i.e. the metric-level provides additional 'fine-grained' information to the road topology: the metric representation can be reduced to the topological representation, but not vice versa. This subsumption architecture was hence used in [17] to demonstrate, in the context of a driver assistance system, a full induction of the intentional hierarchy in human drivers.

We therefore conclude by arguing that Perception-Action learning, as well as enabling autonomous cognitive bootstrapping architectures, also constitutes a particularly straightforward approach to modeling human intentionality, in that it makes fewest cognitive assumptions - the existence of perceptual representation is only be assumed in so far as it directly relates to an observable high-level action concept (such a 'navigating a junction', 'stopping at a red light', etc). This bijectivity of perception and action also gives a natural explanation for wider intention-related phenomenon such as *action mirroring*.

References

[1] H. Dreyfus, *What Computers Can't Do*. New York: Harper and Row, 1972.

[2] C. L. Nehaniv, D. Polani, K. Dautenhahn, R. te Boekhorst, and L. Canamero, "Meaningful information, sensor evolution, and the temporal horizon of embodied organisms," in *Artificial Life VIII*, B. Standish, Abbass, Ed. MIT Press, 2002, pp. 345–349.

[3] J. McCarthy and P. Hayes, "Some philosophical problems from the standpoint of artificial intelligence," *Machine Intelligence*, no. 4, pp. 463–502, 1969.

[4] S. Harnad, "The symbol grounding problem," *Physica D*, no. 42, pp. 335–346, 1990.

[5] G. Granlund, "Organization of architectures for cognitive vision systems," in *Proceedings of Workshop on Cognitive Vision*, Schloss Dagstuhl, Germany, 2003.

[6] M. Felsberg, J. Wiklund, and G. Granlund, "Exploratory learning structures in artificial cognitive systems," *Image and Vision Computing*, vol. 27, no. 11, pp. 1671–1687, 2009.

[7] D. Windridge and J. Kittler, "Perception-action learning as an epistemologically-consistent model for self-updating cognitive representation," in *Brain Inspired Cognitive Systems 2008*. Springer, 2010, pp. 95–134.

[8] ——, "Epistemic constraints on autonomous symbolic representation in natural and artificial agents," in *Studies in Computational Intelligence: Applications of Computational Intelligence in Biology*. Springer Berlin Heidelberg, 2008, vol. 122, pp. 395–422.

[9] D. Windridge, M. Felsberg, and A. Shaukat, "A framework for hierarchical perception-action learning utilizing fuzzy reasoning," *Cybernetics, IEEE Transactions on*, vol. 43, no. 1, pp. 155–169, Feb 2013.

[10] J. G. Wolff, "Cognitive development as optimisation," in *Computational Models of Learning*, L. Bolc, Ed. Heidelberg: Springer-Verlag, 1987, pp. 161–205.

[11] M. Shevchenko, D. Windridge, and J. Kittler, "A linear-complexity reparameterisation strategy for the hierarchical bootstrapping of capabilities within perception–action architectures," *Image and Vision Computing*, vol. 27, no. 11, pp. 1702–1714, 2009.

[12] A. Newell and H. Simon, "The theory of human problem solving; reprinted in collins & smith (eds.)," in *Readings in Cognitive Science, section 1.3.*, 1976.

[13] D. Marr, *Vision: A Computational Approach.* San Fr.: Freeman & Co., 1982.

[14] P. Gärdenfors, "How logic emerges from the dynamics of information," *Logic and Information Flow*, pp. 49–77, 1994.

[15] J. Modayil, "Bootstrap learning a perceptually grounded object ontology," 2005, retr. 9/5/2005 http://www.cs.utexas.edu/users/modayil/modayil-proposal.pdf.

[16] D. L. Hintzman, "Schema abstraction in a multiple-trace memory model," *Psychological review*, vol. 93, no. 4, pp. 411–428, 1986.

[17] D. Windridge, A. Shaukat, and E. Hollnagel, "Characterizing driver intention via hierarchical perception-action modeling," *Human-Machine Systems, IEEE Transactions on*, vol. 43, no. 1, pp. 17–31, 2013.