Emergent Intentionality in Perception-Action Subsumption Hierarchies

David Windridge [1,*]

[1] *Department of Computer Science, Middlesex University, London, UK*

Correspondence*:
David Windridge
d.windridge@mdx.ac.uk

ABSTRACT

A *cognitively-autonomous* artificial agent may be defined as one able to modify both its external world-model *and* the framework by which it represents the world, requiring two simultaneous optimization objectives. This presents deep epistemological issues centered on the question of how a framework for representation (as opposed to the entities it represents) may be objectively validated. In this summary paper, formalizing previous work in this field, it is argued that subsumptive perception-action learning has the capacity to resolve these issues by *a)* building the perceptual hierarchy from the bottom up so as to ground all proposed representations and *b)* maintaining a bijective coupling between proposed percepts and projected action possibilities to ensure empirical falsifiability of these grounded representations. In doing so, we will show that such subsumptive perception-action learners intrinsically incorporate a model for how intentionality emerges from randomized exploratory activity in the form of 'motor babbling'. Moreover, such a model of intentionality also naturally translates into a model for human-computer interfacing that makes minimal assumptions as to cognitive states.

Keywords: Perception-Action Learning, Intention recognition, Embodied cognition, Subsumption Hierarchies, Symbol Grounding

1 INTRODUCTION

Significant deficits have been apparent in traditional approaches to embodied computer vision for some time Dreyfus (1972). In the conventional approach to autonomous robotics, a computer vision system is employed to build a model of the agent's environment *prior* to the act of planning the agent's actions within the modeled domain. Visuo-haptic 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 Settles (2010); Koltchinskii (2010)).

However, it is apparent, in this approach, that there exists a very wide disparity between the visual parameterization of the agent's domain and its action capabilities within it Nehaniv et al. (2002). For instance, the agent's visual parametric freedom will typically encompass the full intensity ranges of the RGB channels of each individual pixel of a camera CCD, such the the range of *possible* images generated per time-frame is of an extremely large order of magnitude, despite the fact that only a minuscule fraction of this representational space would ever be experienced by the agent. (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). On the other hand, the agent's motor capability is likely to be very much more parametrically-constrained (perhaps consisting of the possible Euler angle settings of the various actuator motors). This disparity is manifested in classical problems such as *framing* McCarthy and Hayes (1969) and *symbol grounding*. (The latter occurs when abstractly-manipulated symbolic objects lack an intrinsic connection to the real-world objects that they represent; thus a chess-playing robot typically requires a prior supervised computer vision to be solved in order to apply deduced moves to visually-presented chess pieces.)

Perception-Action (P-A) learning was proposed in order to overcome these issues, adopting as its informal motto, 'action *precedes* perception' Granlund (2003); Felsberg et al. (2009). By this it is meant that, in a fully-formalizable sense, actions are conceptually prior to perceptions; i.e. perceptual capabilities should depend on action-capabilities and not vice versa. (We thus distinguish *PA-learning* from more generalized forms of learning within a perception/action context (cf. e.g. d. R. Millan (2016); Mai et al. (2013); Masuta et al. (2015)), in which the nature of the perceptual domain remains fixed *a priori* [albeit with potential variations in e.g. visual saliency]).

It will be the argument of this paper that perception-action learning, as well as having this capacity to resolve fundamental epistemic questions about emergent representational capacity, also naturally gives a model for emergent intentionality that applies to both human and artificial agents, and may thus be deployed as an effective design-strategy in human-computer interfacing.

2 PERCEPTION-ACTION LEARNING

Perception-Action learning agents thus proceed by randomly sampling their action space ('motor babbling'). For each motor action that produces a discernible perceptual output in the bootstrap representation space S (consisting of e.g. camera pixels), 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*); the agent learns to perceive only that which it can change. More accurately, the agent learns to perceive only that which it *hypothesizes* that it can change - thus, the set of experimental data points $\cup_i p_i \subset S$ can, in theory, be generalized over so as to create an *affordance-manifold* that can be mapped onto the action space via the injective relation $\{actions\} \rightarrow \{percept_{\text{initial}}\} \times \{percept_{\text{final}}\}$ Windridge and Kittler (2010, 2008); Windridge et al. (2013a).

2.1 Subsumptive Perception-Action Learning

Importantly, this approach permits *Cognitive Bootstrapping* Windridge and Kittler (2010), the bootstrapping of an autonomous agent's representational framework simultaneously with the world-model represented in terms of that framework. This centers on the fact that the learned manifold embodying the *injective* relation $\{actions\} \rightarrow \{percept_{\text{initial}}\} \times \{percept_{\text{final}}\}$ represents a constrained subset of the initial action domain, and as such, is 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 $\{actions\} \rightarrow \{percept^{\text{new}}_{\text{initial}}\} \times \{percept^{\text{new}}_{\text{final}}\}$ such that each hypothesizable action (ie intention primitive) has a unique, discriminable outcome Windridge and Kittler (2010, 2008); Windridge et al. (2013a)).

Each induced higher-level action/intention (e.g. *Translate*) is thus created co-extantly with a higher-level percept domain (e.g. *Object*)). The falsifiability of such induced representational concepts arises from actively addressing the question of whether this higher-level perception in fact constitutes a useful description of the world i.e. whether it yields a net compression in the agent's internal representation of its own possible interactions with the world (its affordances). In particular, it is argued in Windridge and Kittler (2008), that the perception-action bijectivity constraint applied in such a hierarchical manner is uniquely sufficient to enable simultaneous empirical falsifiability of the cognitive agent's world model *and* the means by which this world is perceived (by virtue of the implicit grounding of the unique set of higher-level percepts so generated).

Very often parametric 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 Wolff (1987); 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. (Correct in the sense of being a true model of the world's affordances; i.e. every possible instantiation of the induced high-level actions terminates in the anticipated percept, with no possible environmental actions being overlooked.)

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* (Shevchenko et al. (2009)), 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. In the example given, which juxtaposes a simulated camera-equipped robot arm in relation to a child's shape-shorter puzzle, the robotic agent commences by motor babbling in the initial motor-actuator domain, and eventually progresses to motor-babbling in the bootstrapped "move-shape-to-hole" action domain (i.e. placing a randomly chosen object into its corresponding hole). This apparently intentional activity amounts to solving the shape-sorter puzzle, even though the system is still only motor babbling albeit at a higher level of the induced hierarchy, and has no prior programming as to the 'goal' of the environment[1].

Procedurally, this takes place as a recursive loop alternating between *exploration*, *generalization* and *representation* as in Algorithm 1. Note in particular, in Algorithm 1, that the act of parametrically-instantiating a proposed bijective perception-action term $\{P_n^{\text{initial}}\} \times \{P_n^{\text{final}}\}$ with respect to an initial perceptual state $\{P_n^{\text{initial}}\}$ and a sought perceptual end-state $\{P_n^{\text{final}}\}$is equivalent to formulating an *intention* (which may or may not be achievable in the environment).

Algorithm 1 *Ab Initio* Induction of Perception-Action Hierarchy in Artificial Agents

1: **Initialization** Obtain:
2: Bootstrap percept set $\{P_1\}$ (eg camera pixel)
3: Bootstrap action set $\{A_1\}$ (motor primitives)
4: Inference mechanism capable of generalizing exploratory samples from function M
$M : \{P_n^{\text{initial}}\} \times \{P_n^{\text{final}}\} \times \{A\} \rightarrow \{achieved, not\_achieved\}$

5: **while** prediction accuracy $<$ threshold) **do**
6: **A) Carry-out randomized exploratory activity on basis of representational-framework**

i.e. generate grounded top-down parametric instantiations
$A_{i \leq n}(P^{\text{initial}}, P^{\text{final}})$ by randomly selecting initial &
target percepts at *proposed* top level of hierarchy, n

7: **B) Induce rules governing action legitimacy**
legitimate actions achieve intended perceptual goal
($=$ affordance-based model of world)

Generate function $M : \{P\} \times \{P\} \times \{A\} \rightarrow \{true, false\}$
(e.g. via first-order logical induction or stochastic discrimination)

8: **C) Remap perceptual variables to represent novel high-level action hypothesis in most efficient manner**
i.e. form the bijection: $\{A_{n+1}\} \leftrightarrow \{P_{n+1}\} \times \{P_{n+1}\}$
9: **end while**

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[2]. (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. the intention 'stay in the center of the lane', etc). (The hierarchical perception-action paradigm at no stage specifies *how* the scheduled sub-task is to achieve

[1] In this case, the "move-shape-to-hole" action is induced following the failure of the "move-shape-onto-surface" action to produce the anticipated result (i.e. when, following exploratory "move-shape-onto-surface" actions, the object happens by chance to fall into a hole to which matches its shape and orientation). This immediately falsifies the existing rule-base such that that the action domain is necessarily modified, by first-order logical rule-induction, to account for this possibility. In the context of the PA bijectivity condition this requires the existence of an action "move-shape-to-hole" perceptually parameterized by a set of labels corresponding to the perceptual representation of holes. Exploratory instantiation of this higher-level rule then corresponds to placing random objects into their corresponding holes i.e. "solving the shapesorter", even though no such external goal specification has taken place.

[2] The subsumption hierarchy thus acts bidirectionally; the hierarchy is learned bottom-up, while exploratory actions are instantiated top-down.

the perceptual goal - this is free within the framework, and may be achieved by a variety of mechanisms e.g. optimal control, minimum jerk etc).

Moreover, these perceptual goals have no internal content; in a fully-achieved perception-action learning agent, the environment effectively "becomes it *own* representation", Newell and Simon (1976), representing a significant compression of the information that an agent needs to retain. This relates directly to the issue of symbol grounding, a seminal problem in the conceptual underpinning of the classical approach to machine learning Harnad (1990). 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 a system such as first-order logic deduction McCarthy and Hayes (1969)) 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* Marr (1982); Gärdenfors (1994); Modayil (2005); Granlund (2003). They are thus always intrinsically grounded (for an example of utilization of first-order logic induction within a subsumption hierarchy see Windridge and Kittler (2010)).

The subsumption hierarchy is thus typically characterized by continuous stochastic relationships on the lower levels with more discrete, symbolic manipulations occurring at the higher levels - for this reason, consistent with findings of Shevchenko et al. (2009), 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*. (This phenomenon is readily apparent in the development of motor movement of human infants as schema abstraction takes place - for instance, the intuition of a generalized percept category *container* correlates with the attempt to validate this notion via the repeated placing of a variety of objects into a variety of containers; cf Hintzman (1986) for an analysis of scheme abstraction in infants).

Such high-level schema-employment in humans can, in principle, be detected via an appropriate classification system, enabling novel forms of intentional interfacing between humans and machines.

3 HUMAN-COMPUTER INTERFACING

The percept-action relationship may thus 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.

We can thus adopt the perception-action bijectivity principle as a *design paradigm* in building HCI systems by demanding that intentional acts on the part of the user are correlated maximally-efficiently (i.e. bijectively) with perceptual transitions apparent to the user. This thus permits a user interface that makes minimal assumptions as to underlying cognitive processes, assuming nothing more than the ability to discriminate percept termina. This subsumption architecture paradigm was used in Windridge et al. (2013b) to demonstrate, in the context of a driver assistance system, induction of the intentional hierarchy for drivers of a vehicle in which action and eye-gaze take place with respect an external road camera view. The corresponding system constructed for the the project demonstrator was thus able to determine the driver's intentional hierarchy in relation to the current road situation and provide assistance accordingly. In principle, such an interface can also be extended to direct mechanical assistance by substituting the computationally modeled perception-action system for the human perception-action system along the lines of the horse-rider interaction paradigm.

Such P-A HCI interfaces will generally require the ability to adaptively link high-level reasoning processes (modeled by e.g. first-order logic) with low-level reactive processes (modeled, for example, stochastically). This amounts to a requirement to propagate learning across the symbolic/sub-symbolic divide. However, because the P-A hierarchy does not make intrinsic distinction between these (there is only progressively grounded P-A abstraction), it is possible to conceive of P-A learning platforms that embody a variety of different learning approaches at different hierarchical levels, but which are all able to learn together by passing derivatives between hierarchical layers in a manner analogous to deep learning approaches.

An example utilizing a two-layer P-A hierarchy is given in Windridge et al. (2013a) which incorporates a fuzzy first-order logic reasoning process on the top level and an Euler-Lagrange based trajectory optmisation process on the lower level. The fuzzy-reasoning process employs predicates embodying the P-A bijectivity condition to compute the fixed point of the logical operator T_P; i.e. $T_P(I) = I$ for each time interval t.

I is thus the *Herbrand* model, the minimal logically-consistent 'world model' for time t, of the logical programme P (where $P =$*fixed clauses+temporalized detections+ground atom queries for* $t+1$; P hence embodies a series of first-order logical rules concerning traffic behavior). This functionalization of the logical reasoning enables the predicate-prediction disparity with respect to the lower-level to modulate the lower-level's Euler-Lagrange optimization via the inter-level Jacobean derivatives. The net result is logically-weighted updating of the Euler-Lagrange optimization that allows for on-line (top-down and bottom-up) adaptivity to human inputs. For example, in top-down terms, this allows a logically-influenced Bayesian prior for gaze-location at junctions to be derived. It also allows for adaptive symbol tethering; for example actively associating eye-gaze clusters with specific semantically-described road entities (such as stop & give-way signs) via their logical context.

In principle, any high-level abstract reasoning or induction process can be incorporated with low-level stochastic learning in this manner; highly flexible human-computer interfaces are thus made possible through adopting perception-action bijectivity as a *design principle*.

4 CONCLUSION

We have proposed perception-action hierarchies as a natural solution to the problem of representational induction in artificial agents in a manner that maintains empirical validatability. In such *ab initio* P-A hierarchies (i.e. where cognitive representations are bootstrapped in a bottom-up fashion), exploration is conducted via motor-babbling at progressively higher levels of the hierarchy. This necessarily involves the spontaneous scheduling of perceptual goals and sub-goals in the induced lower levels of the hierarchy in such a way that, as the hierarchy becomes deeper, that the randomized exploration becomes increasingly 'intentional' (a phenomenon that is readily apparent in the development of motor movement in human infants).

This has implications for social robotics; in particular, it becomes possible to envisage communicable actions within collections of agents employing P-A hierarchies. Here, the same bijectivity considerations apply to perceptions and actions as before, however the induction and grounding of symbols would be conducted through linguistic exchange (we note in passing that the perception-action bijectivity constraint implicitly embodies the notion of mirroring without requiring specific perceptual apparata - 'mirror neurons' etc).

P-A subsumption hierarchies naturally also encompass symbolic/sub-symbolic integration and permit adaptive learning with respect to existing knowledge bases; in this case a bijective P-A consistency criterion is imposed on the engineered subsumption hierarchy. Moreover, P-A-subsumption hierarchies naturally lend themselves to a "deep" formulation in neural-symbolic terms d'Avila Garcez et al. (2009); this is the subject of ongoing research.

We therefore conclude 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 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); conversely, the ability to correctly interpret a human agent's action implicitly invokes a necessary and sufficient set of perceptual representations on the part of the agent. This bijectivity of perception and action also gives a natural explanation for wider intention-related phenomenon such as *action mirroring*.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

The first author would like to acknowledge financial support from the Horizon 2020 European Research project DREAMS4CARS (no. 731593).

REFERENCES

d. R. Millan, J. (2016). Brain-controlled devices: the perception-action closed loop. In *2016 4th International Winter Conference on Brain-Computer Interface (BCI)*. 1–2. doi:10.1109/IWW-BCI.2016.7457451

d'Avila Garcez, A. S., Lamb, L. C., and Gabbay, D. M. (2009). *Neural-Symbolic Cognitive Reasoning*. Cognitive Technologies (Springer). doi:10.1007/978-3-540-73246-4

Dreyfus, H. (1972). *What Computers Can't Do* (New York: Harper and Row)

Felsberg, M., Wiklund, J., and Granlund, G. (2009). Exploratory learning structures in artificial cognitive systems. *Image and Vision Computing* 27, 1671–1687

Gärdenfors, P. (1994). How logic emerges from the dynamics of information. *Logic and Information Flow* , 49–77

Granlund, G. (2003). Organization of architectures for cognitive vision systems. In *Proceedings of Workshop on Cognitive Vision* (Schloss Dagstuhl, Germany)

Harnad, S. (1990). The symbol grounding problem. *Physica D* , 335–346

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological review* 93, 411–428

Koltchinskii, V. (2010). Rademacher complexities and bounding the excess risk in active learning. *The Journal of Machine Learning Research* 11, 2457–2485

Mai, X., Zhang, X., Jin, Y., Yang, Y., and Zhang, J. (2013). Simple perception-action strategy based on hierarchical temporal memory. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 1759–1764. doi:10.1109/ROBIO.2013.6739722

Marr, D. (1982). *Vision: A Computational Approach* (San Fr.: Freeman & Co.)

Masuta, H., o. Lim, H., Motoyoshi, T., Koyanagi, K., and Oshima, T. (2015). Direct perception and action system for unknown object grasping. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 313–318. doi:10.1109/ROMAN.2015.7333637

McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* , 463–502

Modayil, J. (2005). Bootstrap learning a perceptually grounded object ontology. Retr. 9/5/2005 http://www.cs.utexas.edu/users/modayil/modayil-proposal.pdf

Nehaniv, C. L., Polani, D., Dautenhahn, K., te Boekhorst, R., and Canamero, L. (2002). Meaningful information, sensor evolution, and the temporal horizon of embodied organisms. In *Artificial Life VIII*, ed. B. Standish, Abbass (MIT Press). 345–349

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

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*

Shevchenko, M., Windridge, D., and Kittler, J. (2009). A linear-complexity reparameterisation strategy for the hierarchical bootstrapping of capabilities within perception–action architectures. *Image and Vision Computing* 27, 1702–1714

Windridge, D., Felsberg, M., and Shaukat, A. (2013a). A framework for hierarchical perception-action learning utilizing fuzzy reasoning. *Cybernetics, IEEE Transactions on* 43, 155–169. doi:10.1109/TSMCB.2012.2202109

Windridge, D. and Kittler, J. (2008). 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), vol. 122. 395–422. doi:10.1007/978-3-540-78534-7_16

Windridge, D. and Kittler, J. (2010). Perception-action learning as an epistemologically-consistent model for self-updating cognitive representation. In *Brain Inspired Cognitive Systems 2008* (Springer). 95–134

Windridge, D., Shaukat, A., and Hollnagel, E. (2013b). Characterizing driver intention via hierarchical perception-action modeling. *Human-Machine Systems, IEEE Transactions on* 43, 17–31. doi:10.1109/TSMCA.2012.2216868

Wolff, J. G. (1987). Cognitive development as optimisation. In *Computational Models of Learning*, ed. L. Bolc (Heidelberg: Springer-Verlag), 161–205