Windridge, David (2017) Emergent intentionality in perception-action subsumption hierarchies. Frontiers in Robotics and AI, 4 . ISSN 2296-9144 (Published online first)
http://dx.doi.org/10.3389/frobt.2017.00038
Published version (with publisher's formatting)
Available from Middlesex University's Research Repository at http://eprints.mdx.ac.uk/22394/

Copyright:

Middlesex University Research Repository makes the University's research available electronically.

Copyright and moral rights to this thesis/research project are retained by the author and/or other copyright owners. The work is supplied on the understanding that any use for commercial gain is strictly forbidden. A copy may be downloaded for personal, non-commercial, research or study without prior permission and without charge. Any use of the thesis/research project for private study or research must be properly acknowledged with reference to the work's full bibliographic details.

This thesis/research project may not be reproduced in any format or medium, or extensive quotations taken from it, or its content changed in any way, without first obtaining permission in writing from the copyright holder(s).

If you believe that any material held in the repository infringes copyright law, please contact the Repository Team at Middlesex University via the following email address:
eprints@mdx.ac.uk

The item will be removed from the repository while any claim is being investigated.
Emergent Intentionality in Perception-Action Subsumption Hierarchies

David Windridge*

Department of Computer Science, Middlesex University, London, United Kingdom

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 article, 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 (Koltchinskii, 2010; Settles, 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 that 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...
Cognitive Bootstrapping

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 \( \{\text{actions}\} \rightarrow \{\text{percept}_{\text{initial}}\} \times \{\text{percept}_{\text{final}}\} \) 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 \( \{\text{actions}\} \leftrightarrow \{\text{percept}_{\text{initial}}\} \times \{\text{percept}_{\text{final}}\} \) such that each hypothesizable action (i.e., intention primitive) has a unique, discriminable outcome (Windridge and Kittler, 2008, 2010; Windridge et al., 2013a)).

Each induced higher level action/intention (e.g., Translate) is thus created coextantly 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 \( P \) (with \( P \) and \( n \) being the intensity resolution and number of pixels, respectively), while an object-based representation typically has a parametric magnitude of \( \sim n^2 \), 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 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-sorter 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 \( \{P^{\text{init}}\} \times \{P^{\text{end}}\} \) with respect to an initial perceptual state \( \{P^{\text{init}}\} \) and a sought perceptual end-state \( \{P^{\text{end}}\} \) is equivalent to formulating an intention (which may or may not be achievable in the environment).

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 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.)

³In 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 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 shapeniner," even though no such external goal specification has taken place.

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

---

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

1. Initialization Obtain:
2. Bootstrap percept set \( \{P_i\} \) (eg camera pixel)
3. Bootstrap action set \( \{A_i\} \) (motor primitives)
4. Inference mechanism capable of generalizing exploratory samples from function \( M : \{P^{\text{init}}\} \times \{P^{\text{end}}\} \rightarrow \{\text{achieved, not achieved}\} \)
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 \( A_{\text{ind}}(P^{\text{init}}, P^{\text{end}}) \) 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
   - (w = affordance-based model of world)
   - Generate function \( M : \{P\} \times \{P\} \times \{A\} \rightarrow \{\text{true, false}\} \)
   - (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: \( \{A_{\text{ind}}\} \leftrightarrow \{P_{\text{ind}}\} \times \{P_{\text{ind}}\} \)
9. endwhile

Moreover, these perceptual goals have no internal content; in a fully achieved perception-action learning agent, the environment effectively "becomes its own representation" (Newell and Simon, 1972), 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; Granlund, 2003; Modayil and Kuipers, 2004). 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 subgoals 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 to an external road camera view. The corresponding system constructed for the project demonstrator was thus able to determine the driver’s intentional hierarchy in relation to the current road situation and provide guidance 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 optimization 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., $T_P(t) = I$ 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 $P = \text{fixed clauses} + \text{temporalized detections} + \text{ground atom queries}$ 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 interlevel 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 and 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 subgoals 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 apparatus—“mirror neurons,” etc.).

P-A subsumption hierarchies naturally also encompass symbolic/subsymbolic 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 as “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.

AUTHOR CONTRIBUTIONS

The author is responsible for all aspects of the work.

FUNDING

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

REFERENCES


Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2017 Windridge. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.