Addressing the double contingency problem with approaches that generate such learning would allow more sophisticated subsystem organization because each subsystem could control more sophisticated interactions than a linear control loop. Therefore, we anticipate that addressing the problem of subsystem formation driven by double contingency within the general framework of learning by experiencing would allow more advances in constructivist agent design.

Olivier L. Georgeon is currently an associate researcher at the LIRIS Lab, with a fellowship from the French Government (ANR-RPDOC program). He received a Masters in computer engineering from Ecole Centrale de Marseille in 1988, and a PhD in psychology from the Université de Lyon in 2008.

Received: 19 February 2014
Accepted: 19 February 2014

Aligning Homeostatic and Heterostatic Perspectives

Patrick M. Pilarski
University of Alberta, Canada
pilarski/at/ualberta.ca

> **Upshot** • There is merit to the continuous-signal-space homeostatic viewpoint on subsystem formation presented by Bernd Porr and Paolo Di Prodi; many of their ideas also align well with a heterostatic constructivist perspective, and specifically developments in the field of reinforcement learning. This commentary therefore aims to identify and clarify some of the linkages made by the authors, and highlight ways in which these interdisciplinary connections may be leveraged to enable future progress.

« 1 » Learning to perceive, predict, and act based only on continuous-valued sensorimotor inputs and outputs is a challenging and important pursuit that deserves our focused attention. While subsystem formation and evaluation are the principal listed contributions of Bernd Porr and Paolo Di Prodi's target article, the nature of the signals and predictions in the paper's problem domain are crucial points that impact how we interpret the comparisons made in the paper and the ways its insights may be applied to work in other domains. I use Porr and Di Prodi's problem setting and agent formulation as a starting point for assessing some of the key statements made in their work, and build toward a specific look at prediction utilization as presented by the authors. This assessment is supplemented with comparisons to related work from the recent computational and biological literature.

Heterostasis and homeostasis

« 2 » Porr and Di Prodi's setting of agents interacting in a reflexive and predictive manner via continuous inputs and outputs is a natural one, albeit one that is often ignored in favour of the perceived clarity and mathematical benefits of discrete sensation and action spaces. Their specific setting is in fact a problem domain that resonates well with other robot-related constructivist demonstrations from the machine learning literature – e.g., learned multi-robot food foraging behaviour (Matarić 1997), robot learning applications as surveyed by Grondman et al. (2012), and robot knowledge acquisition as per Modayil, White & Sutton (2014) and Sutton et al. (2011, as cited by the authors). It is important to note, however, that many of these like-minded explorations are rooted in a rather different starting point: that of the learning system or systems attempting to maximize some aspect of its experience – in other words, an agent seeking to increase its long-term expected reward or learning progress, as in the intertwined fields of computational and biological reinforcement learning (Sutton & Barto 1998). This maximization, or *heterostatic* goal-seeking behaviour (after Harry Klopf's *The Hedonistic Neuron*, 1982) is at first glance in contrast with an agent's "task of restoring its desired state to homeostasis," as posed by the authors (§3). However, for our current discussion, it may be beneficial to explore the similarities between these viewpoints in terms of the authors' work, as opposed to the differences.

« 3 » Let us first examine the statements made in the authors' concluding remarks, suggesting that the homeostatic linear control approach in the paper "establishes an ongoing process that has no final goal but is rather driven by intrinsic motivations that are defined by desired states." (§68) After acknowledging the need and potential for more complex actions and action sequences, as potentially provided by techniques from reinforcement learning, the text of §68 continues by stating that the extrinsically defined reward used in standard reinforcement learning "usually means that the life of the animal is just directed toward this single moment in time but will not code an ongoing intrinsic motivation."[1] This sequence of text sets up a natural contrast between extrinsic and intrinsic reward – motivation or satisfaction derived from the world or from within the agent, respectively. At the same time, it reinforces a distinction between heterostatic and homeostatic optimization by an intelligent system.

« 4 » Intrinsic motivation is held to be a powerful way to drive exploration and potentially accelerate the learning of predictions, control behavior, and better representations (Schmidhuber 1991; Oudeyer, Kaplan & Hafner 2007). However, much like the actual boundary between an agent and its environment is often less of a boundary and more an opinion on the part of the system designer (or examiner), boundaries between what are considered intrinsic and extrinsic reward have been placed at different points by different authors. Is the distinction between these types of feedback actually useful to our discussion of the present paper, or does it further cloud the understanding of how Porr and Di Prodi's agents react to perturbations in their sensorimotor streams?

« 5 » One high-level view we could be inclined to take based on the statements made in §68 is that an intrinsic approach to motivation allows ongoing, life-long learning without the need for endpoints or imposed valuations of an agent's stream of experience (e.g., transient or final goals). However, it is interesting to refer again to the aforementioned text in §68 indicating

[1] This statement seems to assume a terminal or discrete reward, and passes over the way that standard reinforcement learning often utilizes temporally extended expectations of future reward (e.g., *discounted future return;* Sutton & Barto 1998) or average reward (discussed below). However, a detailed discussion of all these points is best left outside the present commentary.

that an agent has “desired states.” We also see this allusion to desire, or the relative valuation of changes to an agent’s input, phrased as “irritations” (e.g., §5 and §16). Whether intrinsic or extrinsic, valuation and goal seeking seems to play a role in the authors’ learning system and the way they describe that system.

« 6 » Here again, language may be clouding our view on the learning setting of interest. Let us return to a heterostatic view via reinforcement learning, where goal-based desire and state-action valuation is commonplace. Reinforcement learning is a form of optimal control, where prediction change and policy change are driven by temporal-difference error signals derived from the incoming stream of data. One signal is labeled as special, namely, the reward. This is not so different from the learning presented by the authors, wherein the difference between two sensors on their simulated robot is formulated as an error signal (thought of as a negative outcome, §5 and §16) and used to drive weight change in another unit. This differential learning signal could also be phrased as reward – extrinsic or intrinsic, depending on where an examiner decides to place their boundaries. This suggestion to convert a problem from the homeostatic to heterostatic viewpoint and back again is perhaps not a new one, but it is worth recalling when we think on the potential extensions and impact of Porr and Di Prodi’s paper.

« 7 » Can we then appeal to the continuous, ongoing nature of a control approach or signal space, for example, Ashby’s homeostat (Ashby 1956), as a clear difference between the authors’ outlined constructivist viewpoint and the constructivist viewpoint embodied by computational reinforcement learning? Here again, the distinction is not easy to pin down without invoking specific (and possibly distracting) language. A well-known setting in reinforcement learning is that of *average reward*, as described in Sutton & Barto (1998),[2] Grondman et al. (2012), and specifically for a continuous-action-space control learner in Degris, Pilarski & Sutton (2012). In the average reward setting, best thought of in terms of continuing problems with no clear end or stopping point (e.g., life-long control learning that proceeds from raw sensorimotor data), the error signals that drive an agent’s valuations of the goodness or badness of a situation are calculated in terms of the difference between the instantaneous reward (arriving at every slice of time from either a privileged reward channel or as computed from a combination of the agent’s sensors) and an adaptable baseline that has been acquired through ongoing experience – the state-invariant average reward. In other words, in the average-reward setting, “one neither discounts nor divides experience into distinct episodes with finite returns. […] one seeks to obtain the maximum reward *per time step*” (Sutton & Barto 1998: 153; my emphasis). This can perhaps also be thought of as minimizing the unpleasant divergence from a steady homeostatic state, as posed by the authors as the engagement of reflex actions in §23 and §27.

Prediction-driven behaviour

« 8 » With these possible symmetries in mind, let us turn to Porr and Di Prodi’s use of learned predictions as inputs, and as incrementally engaged motor alternatives to hardcoded reflex actions. This usage is another key point and incredibly valuable perspective put forward by the authors’ work, and it echoes findings in the study of the human brain. In particular, the authors’ experimental formulation links nicely with the way predictions are thought to be incrementally learned (e.g., via the cerebellum) and then mapped in a fixed way to influence or determine animal action selection – i.e., the “Pavlovian action selection” of David Redish (2013) and work by David Linden (2003). Though Porr and Di Prodi’s predictions are perhaps best thought of as long-range sensors that gradually come online, as opposed to predictions that are built up over time, the anticipatory ideas are largely the same. Their approach can be framed as a view into the gradual acquisition and use of *predictive representations of state* (Littman, Sutton & Singh 2002).

« 9 » As one specific example, Porr and Di Prodi’s ideas resonate with our group’s recent demonstration of how predictions that are learned and adapted in real time can help remove control delays – well viewed as negative outcomes as per §5 – during the operation of a physical human-robot interface (Pilarski, Dick & Sutton 2013). By combining continuous-action reinforcement learning control approaches with knowledge acquisition methods from Sutton et al. (2011) and Modayil, White & Sutton (2014), it was shown that a robot learner could use its predictive inputs to actuate joints preemptively. The robot learned to act prior to stumbling upon its motor objective and being forced rapidly to correct its orientation. This example has an interesting symmetry with Porr and Di Prodi’s setting of using predictions to steer toward a food source or other agent preemptively so that the system is not forced to take rapid, reflexive action at a future point. Effecting Porr and Di Prodi’s shift from reflex to anticipatory actuation in the setting described by Pilarski, Dick & Sutton (2013) – either using divergence from homeostasis, reward, or motor commands related to changes in prediction magnitude – would be a nice demonstration and highlights one way that the present paper by Porr and Di Prodi may catalyze development in other constructivist settings.

« 10 » So, with the discussion above as background, could we reasonably phrase Porr and Di Prodi’s experimental setup in heterostatic terms with the learned, temporally extended predictions of Modayil, White & Sutton (2014) and Sutton et al. (2011) in place of distal sensors? First, would there be any benefits to formulating this parallel view and further validating the authors’ observations? Second, if we did so, would the results and subsystems formed by the interacting learners be similar? It stands to reason that they might, and this would be a potentially valuable link. In addition to this link extending the impact of Porr and Di Prodi’s work into a wide and continually growing body of reinforcement learning literature, it may also be possible for a broader community of researchers to leverage modern developments from reinforcement learning that include representation learning, stable off-policy learning of predictions and control, continuous-state-and-action control learning, and planning that is grounded in sensorimotor experience.

[2] A second edition is in progress and can be accessed at http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html; this new edition comprehensively describes the average reward setting.

A thought experiment

« 11 » As a concrete suggestion: we could perform a simple experiment to begin investigating the parallel, heterostatic view on the authors' work. This would help us identify whether the subsystem results presented by the authors might indeed arise if the multiple agents in their simulated domain were instead basing their behavior on reward-driven learning and learned predictive representations of state. One case would be to implement a control learner that takes as input the differential signals from all predictive sensors and reflex sensors, outputs two continuous-valued motor commands (§21), and receives a negative reward proportional to divergence of the reflex sensors from their differential set points (Figure 1a).[3] To provide a starting behavior identical to that specified by Porr and Di Prodi, the learner's control policy should be initialized to generate reflexive actions as in §23–§28. As a further extension and alternative to prespecified distal sensors, each predictive sensor could instead be implemented such that it reports situation-specific, long-range forecasts about the output of a single reflex sensor at one or more time scales (Figure 1b). As noted above, forecasts of this kind can be incrementally acquired during sensorimotor interaction, and can also be acquired even when an agent is not pursuing a behaviour specifically related to the prediction of interest (i.e., *off-policy learning* as in Sutton et al. 2011).

« 12 » Examining the case of two agents and a single food source, one agent would invariably contact the food first (as indicated in §56). In addition to inducing policy change, this interaction would prevent the second agent from experiencing the food and thus building up anticipatory knowledge related to food attraction (temporally extended predictions) and the related motor responses (linked in either a fixed or a learned way to the magnitude of its predictive signals). The second agent would, however, be able to build up anticipations regarding food stealing since the first agent is now carrying food. Building on the initial perturbations experienced by the two (or more) agents, the formation of subsystems could then follow as per Porr and Di Prodi's discussion. This hypothesis can and should be empirically tested for both the simple case of distal sensors (Figure 1a) and the case where predictive sensors are implemented as incremental prediction learners (Figure 1b).

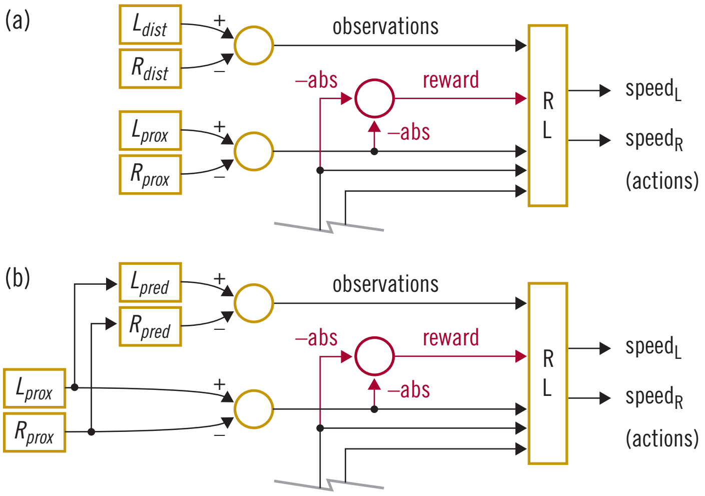

Figure 1: A reinforcement learning formulation for validation experiments using (a) proximal reflexive sensors combined with distal sensors and (b) proximal sensors combined with learned, temporally extended predictions. The generation of differential inputs (observations) to the learning system (denoted RL) is shown for one behaviour type only for clarity; an identical structure can be assumed for the complete system of food attraction and food stealing.

Conclusion

« 13 » There are many good things to take away from Porr and Di Prodi's "Subsystem Formation Driven by Double Contingency." These contributions should be evident to the careful reader, so the aim of this commentary has been to bring out points that may be missed when focusing on the larger claims of the work. This commentary also aimed to emphasize points that may unify our thinking in small ways such that we can move forward more quickly in the understanding of learning and adapting multi-agent systems. Finally, we should ask whether the alignment of homeostatic and heterostatic constructivist viewpoints proposed in this commentary is useful as a default mindset. Likely not. But the quest for life-long, learner-driven representation, prediction, and control could involve a long and bumpy road; I suggest that we need patience and the ability to see both commonalities and differences to follow this road through to its fruitful conclusion.

Patrick M. Pilarski is an Adjunct Assistant Professor in the Department of Computing Science, University of Alberta. As a researcher with the Alberta Innovates Centre for Machine Learning and the Reinforcement Learning and Artificial Intelligence Laboratory, his work focuses on real-time machine learning and adaptive brain-body-machine interfaces for use with assistive robots.

Received: 20 February 2014
Accepted: 24 February 2014

[3] Another plausible example is to examine the drive for agents to return to a desired or expected food carrying state, e.g., positive reward for carrying food (obtained by any means) or negative reward for dropping below an average baseline. This case can also be seen from multiple viewpoints, and might lead to the same type of subsystem formation and reflex-prediction shifts noted by the authors.