Polani, D., (2011). An Informational Perspective on How the Embodiment Can Relieve Cognitive Burden. In *Proc. IEEE Symposium Series in Computational Intelligence 2011 — Symposium on Artificial Life*, 78–85. IEEE.

An Informational Perspective on How the Embodiment Can Relieve Cognitive Burden

Daniel Polani
School of Computer Science, University of Hertfordshire
Hatfield AL10 9AB, Hertfordshire, UK
E-mail: d.polani@herts.ac.uk

***Abstract*—Living organisms are under permanent pressure to take decisions with an impact on their success. Such decisions require information, which can be formulated in the precise sense of Shannon information. Since information processing is costly for organisms, this creates an adaptive pressure for cognition to be as informationally parsimonious as possible. Combining information theory with the theory of reinforcement learning for modeling tasks, we present a number of quantitative analyses how the cognitive burden of an agent deriving from a task can be relieved by the environment and, more specifically, its embodiment. This can be interpreted as moving towards a giving a specific and precise quantitative meaning to Paul's and Pfeifer's concept of *morphological computation* and highlighting the central importance of the embodiment for the success of cognition.**

I. Introduction

One of the goals of the studies of Artificial Life is to identify universal principles governing the dynamics of organisms, which are not tied to a particular substrate and which abstract away the particular biological "implementation", and thus carry over to artificial agents. Various approaches, such as dynamical systems modeling [1], cellular automata [2], [3] and many others have been suggested for this purpose. In the last decade, a new class of approaches based on information theory has been receiving increasing attention. In contrast to the former, it aims at addressing aspects of Artificial Life in a *mechanism-free* way: instead of aiming to specify mechanisms modeling particular phenomena, one specifies (e.g. optimality) principles which result in the desired phenomenon, without making any assumptions about which mechanisms would actually "implement" these principles. This allows for different choices of mechanisms as long as they result in the same macrodynamics. It provides a mesoscopic level of modeling between high-level, phenomenological and low-level, fine-grained models of Artificial Life.

II. Information and Cognitive Processing

Interestingly, information theory has already been recognized as an important potential tool for cognitive modeling only shortly after the concept of information had been introduced in [4], namely in the context of cybernetics and biology [5]–[7]. Evidence for information-maximization principles in biology [8]–[10] and for sensors operating at the physically possible limits of information acquisition [11]–[14] indicates that informational optimality is a candidate for a principle of central importance to biological organisms.

To use such principles, one needs to cast an agent operating in its environment as a control scenario, in which an agent interacts with the environment exerting a certain amount of control over it [15]–[18]. The informational picture of the perception-action loop has studied in various contexts and scenarios [19]–[22]. More importantly, however, these considerations allow one to formulate fundamental limits on the minimal amount of information required for particular tasks, be it a reduction of environmental entropy [15] or the navigation to a target position from a random starting position [18].

Whereas in Shannon's original scenario there is no mechanism to formulate a particular "semantics" or purpose of how transmitted information is to be actually used the *information bottleneck method* demonstrates how to separate relevant from irrelevant portions of Shannon information [23]; this can be extended towards agent scenarios by "qualifying" information via a utility function which attaches a value to each action an agent takes in a particular state [24]. In the case of rewards delayed over a prolonged period, this utility can be modeled by so-called Markovian Decision Processes, MDPs (which are studied in Reinforcement Learning), and combined with the information-theoretic view [25]–[27].

The bottom line of these considerations is that, for an agent to take a decision that achieves particular utility in its world, a certain minimum amount of information processing is *necessarily* required. Together with aforementioned hypothesis that information is costly for organisms, this suggests that organisms would obey a principle of *information parsimony*, minimizing the information required to achieve a sufficient utility [14].

Assuming that the information processing cost gives a quantitative characterization of the "cognitive load" of an agent, we are going to study how this cognitive load can be partly relieved by the environment. It will turn out that, under this view, not just the structure and dynamics of the environment *per se* is important, but how it relates to the particular task, and, more specifically, also the *embodiment* of the agent which we will here intend to mean how precisely the agent is linked into the environment[1] This suggests a quantitative interpretation of the phenomenon of "morphological" or "environmental" computation which has been postulated as basis for the success of suitably embodied agents [28]–[30].

The paper is structured as follows: in Sec. III we introduce basic notation, notions and principles of the MDPs models used, in Sec. IV we explain how these are expanded towards the informational framework, in Sec. V we present the experiments and results, concluding in Sec. VI with a discussion.

III. Model

In the following, we present the model from [25] which will be used for the experiments. The agent's preferences and decision process is modeled as a Markovian Decision Process (MDP). MDPs are a popular approach for modeling sequences of decisions taken by an agent in the face of delayed accumulation of rewards. The structure of the rewards defines the tasks the agent is supposed to achieve. In the present paper, we will restrict ourselves to a simple navigation task, but the formalism is significantly more general [31]. We begin by introducing general notation and the MDP concept.

A. Notation and Definitions

1) Probabilities: First, we introduce some notation and conventions. We use uppercase characters $X, Y, Z, \dots$ for random variables, lowercase characters $x, y, z, \dots$ for the values they assume and curved characters $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \dots$ for their respective domains which we always assume finite. The probability that a random variable X assumes a value $x \in \mathcal{X}$ is written $P(X = x)$, however for simplicity we will use the simpler form $p(x)$ instead by abuse of notation whenever there is no danger of confusion. In particular, in writing $p(x)$ we will not make an explicit distinction between the distribution of the random variable X and the probability value $p(x)$ for the particular outcome $x \in \mathcal{X}$. We write $p(x, y, z)$ for the joint distribution of random variables X, Y, Z, and $p(y|x)$ for the conditional distribution of Y given X.

2) Entropy and Mutual Information: Given a random variable X, define its *entropy* $H(X)$ as $H(X) := -\sum_{x\in\mathcal{X}} p(x) \log p(x)$ where we always assume a binary logarithm. Thus the entropy will be expressed in *bits*, and is a measure of the uncertainty about the outcome of the random experiment X. For jointly distributed variables X, Y, the entropy is defined as $H(X, Y) := -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x, y) \log p(x, y)$ which is equivalent to the entropy of the (single) joint random variable (X, Y). For the random variable pair X, Y, the conditional entropy is defined as

$$\begin{aligned} H(Y|X) &:= \sum_{x\in\mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \ . \end{aligned}$$

Finally, define the *mutual information* between X and Y as

$$I(X; Y) := H(Y) - H(Y|X) \ , \tag{1}$$

i.e. the reduction in uncertainty about the outcome of Y if the outcome of X is known.

3) Markovian Decision Processes (MDPs): Informally, an MDP is a model for an agent taking sequential decisions in an environment with the following properties:

1) The world consists of states and an agent which has a *policy* that determines which actions it selects in which states.
2) After each action taken, the agent obtains a reward (which may be negative). These rewards are cumulated over the lifetime of the agent and determine its achieved utility.
3) Being Markovian, MDPs have no hidden states — that means that, in principle, the agent has full access to the state of the world. We will discuss this assumption briefly in Sec. IV-C. For the particular study, this is not a restriction.

We now formally define MDPs, adopting in the notation from [31] with slight modifications. A *Markovian Decision Process* is defined by its set of states $\mathcal{S}$, its set of actions $\mathcal{A}$, and the pair $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$ defined for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$; here $\mathbf{P}^{s'}_{s,a}$ is the probability that by performing an action a in a state s, the agent will move to state s' and $\mathbf{R}^{s'}_{s,a}$ is the expected reward for this particular transition.

$\mathbf{P}^{s'}_{s,a}$ defines a transition ("structure of the world") and $\mathbf{R}^{s'}_{s,a}$ a reward structure. Given $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$, an agent can employ a *policy* π which specifies its decision process: an action a in a state s is selected with

[1] Here, we will, unlike some other work, not make a distinction between "real" and "simulated" scenarios in using the term *embodiment*.

probability $\pi(a|s)$. Over the course of a single run, an agent will accumulate a reward $\sum_{t'=t}^{\infty} r_{t'}$, starting at time[2] t. The expectation value for this cumulated reward is obtained by averaging over the transition probabilities $\mathbf{P}_{s,a}^{s'}$ and the policy $\pi(a|s)$. Given a starting state s and a policy π, this is the *value* $V^{\pi}(s)$ of the state s following policy π. It can be expressed via the recursive Bellman equation

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \cdot \sum_{s' \in \mathcal{S}} \mathbf{P}_{s,a}^{s'} \cdot \left[\mathbf{R}_{s,a}^{s'} + V^{\pi}(s') \right] . \tag{2}$$

This equation can be used as a fixed point iteration (*value iteration*) for V^{π} by inserting an estimate for V^{π} on the right side and obtaining an improved estimate for it on the left side until convergence. Sometimes it is convenient to further decompose this equation into the Q^{π} function which distinguishes the values attained for a given state s as different actions a are applied:

$$Q^{\pi}(s,a) = \sum_{s' \in \mathcal{S}} \mathbf{P}_{s,a}^{s'} \cdot \left[\mathbf{R}_{s,a}^{s'} + V^{\pi}(s') \right] . \tag{3}$$

$Q^{\pi}(s,a)$ is the utility attained if, in state s, the agent carries out action a, and after that begins to follow π. This representation of the utility allows one to directly evaluate different actions a in a given state s.

In the traditional MDP optimization one now seeks a policy π^* which maximizes the — unique — value function $V^*(s)$ for all states s. Here, however, we will be interested in a modification of the problem, incorporating the decision costs into the problem.

4) Notes on the MDP Definition: Before we proceed to do so, we mention the conventions used here. First, we assume transition probabilities $\mathbf{P}_{s,a}^{s'}$ into states which are not successors of s to be 0. Furthermore, here we only consider navigation tasks and model the goal states of our experimental scenarios as absorbing states which the agent cannot leave once reached.

Second, in the traditional MDP definitions, one assumes that, in the most general case, different states s may have different action sets $\mathcal{A}_s$. Here we deviate slightly from this in that we require the action set $\mathcal{A}$ to be the *same* over all states s. The rationale for this requirement is at the core our interpretation of the decision maker as agent: the embodiment of the agent implies a consistent set of "atomic" actions available to the agent throughout the world — an embodied agent always "takes its actions with it" and the available set of action choices from which the agent selects does not change from state to state [3]. The effect of actions, however, will in general differ in various ways from state to state.

This is, from the point of MDPs, a seemingly minor technical requirement which can be easily accomodated[4] and has no tangible consequences. However, this assumption about the embodiment, i.e. about a particular consistent action set available to the agent throughout the world, will turn out in Sec. V-B to have *major* consequences, once we take the information costs of decision making into consideration.

IV. Information in the Decision Process

A. Overview and Rationale

In this paper, we are not concerned with the cost of learning policies, but delegate that consideration to a generic evolutionary or otherwise adaptive "black box" algorithm (concretely, the algorithm given in Sec. IV-D) which computes the policies. The criterion that we will apply instead is that the policy will be informationally parsimonious. We will make this notion precise in the current section.

First, some general qualitative considerations: if there exists only one optimal policy for the MDP, then that policy is unambiguous and has a given information processing cost. However, if there are multiple optimal policies, then asking for the informationally cheapest one among these optimal policies becomes a more interesting question. Even more interesting becomes the issue when we do no longer demand that the solution be perfectly optimal. After all, strict optimality in one criterium is not the typical situation in biologically relevant scenarios, as many other considerations come into play. Thus, if we only require the expected reward $\mathbf{E}[V(S)]$ to achieve a "sufficiently" large value, the information cost for such a suboptimal (but informationally parsimonious) policy will be generally lower. The extreme case is that of a "blind" agent without information processing cost: it follows the same (but possibly probabilistic) policy independently of the state it is in. In the following, we now make these notions precise and reiterate the methods to compute these policies, where we follow the method from [25].

B. Core Model

Consider an MDP $(\mathbf{P}_{s,a}^{s'}, \mathbf{R}_{s,a}^{s'})$ (state set $\mathcal{S}$ and action set $\mathcal{A}$, as in Sec. III-A3). One can consider an agent graphically as a Bayesian Network, Fig. 1. The random variables $S_0, S_1, S_2, \ldots$ denote the (complete) state of

[2] For the sake of simplicity, we do not consider a discount over time.

[3] We exclude actuator evolution or meta-actions, such as the options model.

[4] To model in our notation action sets $\mathcal{A}_s$ that change between different states s, we characterize illegal (unavailable) actions, i.e. actions outside of $\mathcal{A}_s$, by penalizing them via a infinitely negative reward $\mathbf{R}_{s,a}^{s'} := -\infty$.

the world at times $t = 0, 1, 2, \ldots$. Depending on S_t, the action A_t at each time t is selected according to the policy $\pi(a_t|s_t)$ which is fixed throughout the run. Depending on the particular given state s_t and the action selected a_t, the new state is generated according to probability distribution $p(s_{t+1}|s_t, a_t) \equiv \mathbf{P}^{s_{t+1}}_{s_t,a_t}$.

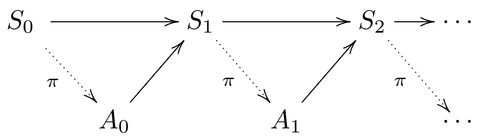

Fig. 1. Bayesian Network indicating the decision structure for an agent

The decision cost incurred by the agent is given by the mutual information

$$I(S;A) = \sum_{s\in\mathcal{S}} p(s) \sum_{a\in\mathcal{A}} \pi(a|s)\, \log \frac{\pi(a|s)}{\sum_{s'\in\mathcal{S}} \pi(a|s')\, p(s')} \tag{4}$$

(essentially a reformulation of (1)). Note that we consider (4) independent of time: we will assume a "steady state" where we do not start the decision process at a specific time but "tap" randomly into the decision process. We will thus simplify the discussion by assuming a fixed distribution of the states $p(s)$ for all time steps.

With these assumptions, the information cost given by (4) then depends only on the state distribution $p(s)$ and the policy π and, since we here consider $p(s)$ as fixed, our only variable of interest becomes π. We now proceed to determine a policy π which is informationally parsimonious, i.e. which minimizes $I(S;A)$ for a given utility level $\mathbf{E}[V^\pi(S)]$.

C. Informationally Optimal Policies

For didactic reasons, we describe the issue of informationally parsimonious solutions in two steps. We first consider only optimal strategies which achieve the (unique) optimal value function $V^*(s)$. This is only achieved if a (not necessarily unique) optimal policy π^* is used by the agent. If the optimal policy is not unique, then one can impose an additional optimality principle amongst the optimal policies, namely seeking one that is informationally parsimonious; i.e. one seeks an optimal policy π^* such that, in addition, $I(S;A)$ as given by (4), with π^* substituted for π, becomes minimal. Such a policy $\bar{\pi}^*$ is called informationally optimal.

An informationally optimal policy $\bar{\pi}^*$ can be interpreted in various ways:

1) among the optimal policies, it requires the least amount of (Shannon) information to distinguish the states S the agent is in;
2) this can be interpreted as the strongest restriction (in terms of information) of the MDP to a process where the state can only be partially observed (a *partially observable MDP*) but where still an optimal value can be reached without the use of memory (see also [27]);
3) alternatively, this can be interpreted as the minimal cost on sensory processing power of a memoryless agent achieving an optimal policy.

To compute informationally optimal policies, one first determines the optimal value function $V^*(s)$ in one of the well-established ways (e.g. by alternating value iteration and then selecting greedy policies, Sec. III-A3) [31]. With the optimal value function $V^*(s)$ one then uses the Lagrangian formalism to formulate the unconstrained minimization problem

$$\min_\pi \Big(I(S;A) - \beta \cdot \mathbf{E}[Q^*(S,A)] \Big) \tag{5}$$

for infinite (in practice very large) β where the expectation $\mathbf{E}$ is taken over the joint distribution of states S and actions A given by $p(s,a) = \pi(a|s)p(s)$. This turns out to be virtually identical with the so-called rate-distortion problem from information theory [23], [32], for which the Blahut-Arimoto fixed point iteration is well established. It consists of a double iteration alternating updates for the policy π and the resulting action distribution $p(a) = \sum_s \pi(a|s)p(s)$ to compute an informationally optimal policy $\bar{\pi}^*$:

$$\pi^{(k)}(a|s) = Z^{-1} \cdot p^{(k-1)}(a) \cdot \exp\left(\beta\, Q^*(s,a)\right) \tag{6}$$

$$p^{(k)}(a) = \sum_{s\in\mathcal{S}} \pi^{(k-1)}(a|s) \cdot p(s) \tag{7}$$

where $\pi^{(k)}$ and $p^{(k)}(a)$ are the estimates for policy and action distribution in the k-th iteration step and Z is a normalization factor. Under mild conditions, this iteration converges to a solution for (5). As in [25], we call the resulting mutual information $I(S;A)$ for a value-wise and informationally optimal policy $\bar{\pi}^*$ *relevant information* for the given MDP.

D. Informationally Suboptimal Policies

We are now introduce the general methodology for suboptimal policies, policies that achieve a particular, but no longer optimal, value $\mathbf{E}[V^\pi(S)]$. In Sec. IV-C, 1where we considered optimal policies only, we computed first the optimal value function $V^*(s)$ and from it, via (3), $Q^*(s,a)$. This optimal value which does not depend on the policy and is universal for an MDP scenario was used in the iterations (6),(7). This is no longer true, however, when we seek policies π that are

Polani, D., (2011). An Informational Perspective on How the Embodiment Can Relieve Cognitive Burden. In *Proc. IEEE Symposium Series in Computational Intelligence 2011 — Symposium on Artificial Life*, 78–85. IEEE.

informationally optimal at a suboptimal value level, since in these cases, the value function $V^\pi(s)$ and its associated utility $Q^\pi(s,a)$ will in general depend on the policy.

While we still can write the Lagrangian minimization task as

$$\min_\pi \Big(I(S;A) - \beta \cdot \mathbf{E}[Q^\pi(S,A)] \Big) \ , \qquad (8)$$

now not only $I(S;A)$, but also $Q^\pi(s,a)$ depends on the policy π. Thus a solution for (8) must be self-consistent not only with respect to (6),(7), but also with respect to the Bellman equation (2). This double self-consistency criterium can be used to derive an algorithm for finding solutions for (8): a single step of value iteration (2) is followed by a single step of the Blahut-Arimoto update (6),(7), repeating until convergence. This method was proposed in [25], and the universal convergence of an extension of this algorithm been conjectured [27], but not proven.

The computations in the following will all use this algorithm[5]. For $\beta \to \infty$, the algorithm computes an optimal strategy that is also informationally optimal, consistent with (5) which uses the optimal Q^* directly. For smaller β, the algorithm produces policies π that are informationally optimal for a given value $\mathbf{E}[V^\pi(S)]$. The trade-off curves in Fig. 3 can be read in two ways: either as the least information $I(S;A)$ to reach a particular value $\mathbf{E}[V^\pi(S)]$ or the best value $\mathbf{E}[V^\pi(S)]$ that can be reached for a fixed information $I(S;A)$. In particular, in the latter, $\mathbf{E}[V^\pi(S)]$ will increase monotonically with growing $I(S;A)$.

V. Experiments

We consider two main scenarios. In both scenarios, we use a square grid world with a varying goal for each of the scenarios. An agent is located in a cell of the grid world, and can take one of four actions, moving it `north`, `east`, `south` or `west` from the cell the agent is currently in.

To implicitly specify the goal position in the scenarios, we define the reward structure $\mathbf{R}^{s'}_{s,a}$ as follows: for each step taken outside the goal state, the reward is -1 (penalty). The grid world is finite and its boundaries are delimited by "walls"; if an action moves the agent into the walls, the agent does not move, but incurs the usual reward -1. Once the goal is reached, the agent does not move away from it and all further rewards are 0 — the task has finished. The value function $V^\pi(s)$ gives the negative of the expected duration of the travel from a given state s to the given goal if the agent follows policy π.

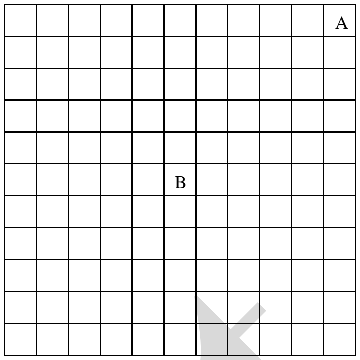

Fig. 2. Grid world with goal positions.

A. Value-Information Trade-Offs for Goal Variations

We now specify the scenarios in detail. Consider a 11×11 square (Fig. 2). Here, we consider two cases: a goal at the top right corner of the grid (A) and a goal in the center of the grid (B). Assuming that the start position is equidistributed over the grid, case B has the shorter average shortest path lengths to the goal and thus the higher optimal values $V^*(s)$ since this value is the negative of the path length.

Figure 3 shows the trade-off between the value achieved for given information $I(S;A)$ for cases A and B under the self-consistent condition (8) for $\beta \in (0,\infty)$. The top right corner in each graph corresponds to $\beta \to \infty$, the optimal value $\mathbf{E}[V^*(S)]$ (shortest path to the goal) and the minimum information required to achieve it. As one reduces β, $I(S;A)$ drops and the policy uses less information about the state S, thus leading to a drop of $\mathbf{E}[V^\pi(S)]$. The limit case is where trade-off curve meets the y-axis and the information $I(S;A)$ becomes 0, the curve intersecting with the vertical axis at the best value that can be achieved by an completely blind agent. In case A (solid curve in Fig. 3), the optimal policy requires a relevant information of ≈ 0.166 bit, achieving a value of ≈ -10.1. The other extreme case of a blind agent with $I(S;A) \to 0$, achieved by a policy which selects a `north` or `east` move with probability 0.5 each, still reaches a value of ≈ -14.5, and is thus reasonably effective in reaching the goal.

In case B, the optimal strategy achieves a better value of -5.5, since the target state is in the center of the square. However, compared A, this comes at the price of a considerably higher amount of relevant information, namely of $I(S;A) \approx 1.17$ bit per step

[5] In the scenarios we are considering, we estimated or calculated solutions for the extreme cases $\beta \to \infty$ and $\beta \to 0$ as "sanity check" benchmarks. For intermediate values of β, we have grounds to believe that the algorithm converged to the actual optimal solution; the results are plausible and consistent with the confirmed limit cases; further work is aiming to validate this assumption.

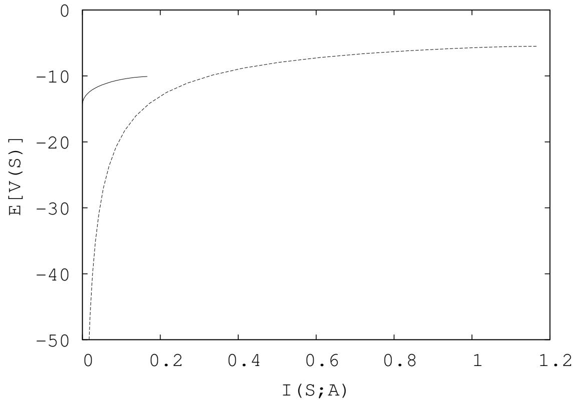

Fig. 3. Trade-off between average value and minimal required information $I(S;A)$ for the square grid MDP from Fig. 2. The horizontal axis shows the information $I(S;A)$ per decision and the vertical axis shows the corresponding value achieved with this information. For higher values, more information is required. The trade-off for A, solid curve, achieves a slightly lower optimal value, but, for most of its parts is strongly favoured to the B trade-off curve.

taken. This effect becomes even more pronounced when one moves on the trade-off curve towards the limit of blind agents $I(S;A) \to 0$. The value in this case goes towards ≈ -205 (outside of the Figure) and the corresponding strategy becomes a purely random walk. The trade-off curve (dashed curve in Fig. 3) for B lies mostly to the right and below that for A. This indicates that case B is, for most of the part far less favourable than case A in terms of "value for information". We will return to these observations.

B. Value-Information Trade-Offs for Relabeled Actions

In the scenarios of Sec. V-A, we varied the goal state between the corner of the world and the centre. In the scenario of the present section, however, we are going to investigate another effect.

1) Action Relabeling: Qualitative Description: In the scenario from Sec. V the agent performed actions `north`, `east`, `south` or `west` which we implied to have the usual effect in the grid world. However, there is nothing in the formalism of MDPs that requires "`north`", "`east`", "`south`" or "`west`" to "mean" the same operation in each state: these are merely labels of four actions that are available to the agent in each state of the world. Intuitively, when we consider an agent "embodied" even in a grid world, we mean action `north` to effect roughly the same operation in each state (with exception of the wall and the goal). However, the MDP formalism allows us to take a "platonic" stance and to assume that the four action directions are just arbitrary labels attached to the actions available to the agent in the current state, with no discernible consistency over different states. More precisely: consider two scenarios, one is the original case A, with the goal in the corner, and the actions labeling the directions of movement in the traditional way. In the second, however, keep the world unchanged, but rename the labels for the four actions in each grid state randomly `north`, `east`, `south` and `west`. This random relabeling is done before the learning run is carried out, the world remains fully deterministic for the agent; the only change is that there is no consistency in the action labels throughout the grid. In other words, the agent does no longer "carry its actions with it".

2) Action Relabeling: Formal Description: We now describe formally the relabeling. This third case, case $\tilde{A}$, consists of a *relabeling* of the actions of case A in the following sense: given an MDP $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$, a *relabeling* of this MDP is given by $(\widetilde{\mathbf{P}}^{s'}_{s,a}, \widetilde{\mathbf{R}}^{s'}_{s,a})$ such that $\widetilde{\mathbf{P}}^{s'}_{s,a} := \mathbf{P}^{s'}_{s,\sigma_s(a)}$ and analogous for $\widetilde{\mathbf{R}}^{s'}_{s,a}$, where σ_s is a permutation $\mathcal{A} \to \mathcal{A}$ of the actions which is, in general, *different* for each $s \in \mathcal{S}$. In the special case of σ_s being the identity permutation for all states $s \in \mathcal{S}$, we reobtain the original MDP.

Importantly, from the point of pure MDP optimization, any relabeling of actions is completely irrelevant. Optimal policies can be computed with the usual value iteration (2), and the resulting values are independent of the relabeling, i.e. one has $V^*(s) = \widetilde{V}^*(s)$ for all $s \in \mathcal{S}$ if $\widetilde{V}^*$ is the optimal value function for the relabeled MDP. More generally, if we operate with a general policy π and consider the Q-function, the Q-values of the original MDP can be related to the new one via the transformation $\widetilde{Q}^{\pi}(s,a) = Q^{\tilde{\pi}}(s, \sigma_s(a))$ where $\pi(s,a) = \tilde{\pi}(s, \sigma_s(a))$, that is $\tilde{\pi}(s,a') = \pi(s, \sigma_s^{-1}(a'))$. In other words, with the exception of an appropriate relabeling of the actions in each state s for a given policy, the relabeled MDP is precisely equivalent to the original one. This is a "platonic" view of the traditional MDP picture: no matter how the "embodiment" (in form of action labels) is modified, it has no consequences for solving the task.

3) Informational Consequences of Action Relabeling: However, once we include the information processing cost into the consideration, this changes drastically. Figure 4 shows again the earlier trade-off curve (solid line) between value and information for case A where actions `north`, `east`, `south`, `west` correspond to the usual directions; furthermore, it shows the trade-off curve for case $\tilde{A}$, where the actions have been relabeled for each state with a different random but fixed permutation (dashed line).

The optimal value for case A had been ≈ -10.1 (Sec. V-A), and this is also the optimal value achieved

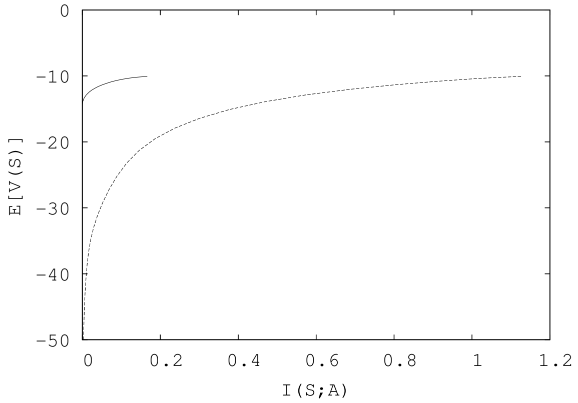

Fig. 4. Value/Relevant Information trade-offs for the square grid MDP. The horizontal axis shows the relevant information per decision and the vertical axis shows the corresponding value achieved with this relevant information. Note that the optimal value achieved is the same for case A (solid line) and case $\tilde{A}$ (dashed line), but at a much higher information cost for case $\tilde{A}$ and generally that curves lies below and right, thus unfavourably to the trade-off curve for case A.

in the relabeled scenario of case $\tilde{A}$, consistently with the discussion in Sec. V-B2.

Now, the differences: while the optimal value $\mathbf{E}[V^*(S)]$ achieved (top-right positions of both curves) is exactly the same for both scenarios, the trade-off curve for the randomly relabeled MDP ($\tilde{A}$) lies far below and to the right to that for A. In particular, for the optimal policy, $\tilde{A}$ requires more than 1.1 bit of information per step. In other words, for the same performance, much more information intake is required in case $\tilde{A}$. And note that, without seeing the action relabeling, an external observer just watching the agent from outside would not reveal any strategy different from that of A. When one now moves towards vanishing $I(S;A)$, i.e. the blind agent, the value drops rapidly[6] to below -222, performing at around the level of for the blind agent of case B.

VI. Discussion

The results show that minor changes in an MDP can induce drastically different outcomes in the informational "metabolism" of the agent. In cases A and B it is clear that the goal in the center is reachable by a slightly shorter average shortest path length than the goal in the corner. However, B requires significantly more information to achieve the optimal solution than A. The effect becomes more pronounced when we reduce information bandwidth; in this case, the achievable value for B drops off very rapidly as compared to A. For A, the wall boundary of the grid helps the agent find the goal in the corner. Even blindly, the agent can randomly select `north` and `east` actions, and the walls will guide it as a funnel towards the goal. For the goal in the center, however, the environment can no longer support the agent in finding the goal: here, a blind agent cannot hope to do better than a random walk.

The role of embodiment in relieving the agent's cognitive burden becomes even more striking in case $\tilde{A}$. All that is dropped from A to $\tilde{A}$ is the consistency of actions ("directions") over the states. From an MDP point of view these are exactly equivalent cases. However, once the cognitive burden is included into the consideration, $\tilde{A}$ is informationally disadvantaged to A. Not only does the optimal case $\beta \to \infty$ require significantly more information per step for $\tilde{A}$, but also, once one moves towards a blind agent, it performs no better than B. Although still in the corner, unlike in A, in $\tilde{A}$ the goal cannot be longer found by the increasingly blinded agent using the wall as "funnel". Instead, the agent needs significantly more information about the current state to identify which two actions in the given state would correspond to the `north`/`east` actions of the original case A. This requires a much larger information intake in $\tilde{A}$, finally leading to the completely uninformed random walk for the fully blinded agent. All that distinguishes case A and $\tilde{A}$ is how the selected action is carried out in the agent's environment.

Concludingly, this provides a prime illustration of the principle of environmental, and more specifically of embodied computation in how embodiment, even in the abstracted view adopted in the present paper, can affect the performance of an agent, once the cognitive burden is taken into consideration,

References:

[1] R. D. Beer, "Dynamical approaches to cognitive science," *Trends in Cognitive Sciences*, vol. 4, no. 3, pp. 91–99, 2000.

[2] C. G. Langton, "Self-reproduction in cellular automata," *Physica D: Nonlinear Phenomena*, vol. 10, no. 1-2, pp. 135 – 144, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/B6TVK-46JH274-7C/2/fc229f72c77e811d6ced201221e2793a

[3] N. Oros and C. L. Nehaniv, "Sexyloop: Self-reproduction, evolution and sex in cellular automata," in *Proc. First IEEE Symposium on Artificial Life (April 1-5, 2007, Hawaii, USA)*, 2007, pp. 130–138.

[4] C. E. Shannon, "The mathematical theory of communication," in *The Mathematical Theory of Communication*, C. E. Shannon and W. Weaver, Eds. Urbana: The University of Illinois Press, 1949.

[6] The value/information trade-off curve at vanishing $I(S;A)$ is almost vertical; the smallest β value in our experiments $\beta = 10^{-4}$, reaching $I(S;A) = 2.4 \cdot 10^{-5}$bit.

[5] W. R. Ashby, *An Introduction to Cybernetics*. Chapman & Hall Ltd., 1956.

[6] F. Attneave, "Informational aspects of visual perception," *Psychol. Rev.*, vol. 61, pp. 183–193, 1954.

[7] H. B. Barlow, "Possible principles underlying the transformations of sensory messages," in *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, W. A. Rosenblith, Ed. The M.I.T. Press, 1959, pp. 217–234.

[8] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, "The metabolic cost of neural information," *Nature Neuroscience*, vol. 1, no. 1, pp. 36–41, 1998.

[9] S. B. Laughlin, "Energy as a constraint on the coding and processing of sensory information," *Current Opinion in Neurobiology*, vol. 11, pp. 475–480, 2001.

[10] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes*, ser. A Bradford Book. MIT Press, 1999.

[11] W. Denk and W. W. Webb, "Thermal-noise-limited transduction observed in mechanosensory receptors of the inner ear," *Phys. Rev. Lett.*, vol. 63, no. 2, pp. 207–210, Jul 1989.

[12] S. Hecht, S. Schlaer, and M. Pirenne, "Energy, quanta and vision," *Journal of the Optical Society of America*, vol. 38, pp. 196–208, 1942.

[13] D. Baylor, T. Lamb, and K. Yau, "Response of retinal rods to single photons," *Journal of Physiology, London*, vol. 288, pp. 613–634, 1979.

[14] D. Polani, "Information: Currency of life?" *HFSP Journal*, vol. 3, no. 5, pp. 307–316, 2009. [Online]. Available: http://link.aip.org/link/?HFS/3/307/1

[15] H. Touchette and S. Lloyd, "Information-theoretic limits of control," *Phys. Rev. Lett.*, vol. 84, p. 1156, 2000.

[16] ——, "Information-theoretic approach to the study of control systems," *Physica A*, vol. 331, pp. 140–172, 2004.

[17] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Organization of the information flow in the perception-action loop of evolved agents," in *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*. IEEE Computer Society, 2004, pp. 177–180.

[18] A. Klyubin, D. Polani, and C. Nehaniv, "Representations of space and time in the maximization of information flow in the perception-action loop," *Neural Computation*, vol. 19, no. 9, pp. 2387–2432, 2007.

[19] O. Sporns and M. Lungarella, "Evolving coordinated behavior by maximizing information structure," in *Proc. Artificial Life X*, L. M. Rocha, M. Bedau, D. Floreano, R. Goldstone, A. Vespignani, and L. Yaeger, Eds., August 2006, pp. 323–329.

[20] M. Lungarella and O. Sporns, "Mapping information flow in sensorimotor networks," *PLoS Computational Biology*, vol. 2, no. 10, 2006.

[21] N. Ay, N. Bertschinger, R. Der, F. Gttler, and E. Olbrich, "Predictive information and explorative behavior of autonomous robots," *European Journal of Physics B*, vol. 63, pp. 329–339, 2008.

[22] K. Zahedi, N. Ay, and R. Der, "Higher coordination with less control — a result of information maximization in the sensorimotor loop," *Adaptive Behaviours*, vol. 18, no. 3-4, pp. 338–355, June 2010.

[23] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annual Allerton Conference on Communication, Control and Computing, Illinois*, Urbana-Champaign, 1999.

[24] D. Polani, T. Martinetz, and J. Kim, "An information-theoretic approach for the quantification of relevance," in *Advances in Artificial Life (Proc. 6th European Conference on Artificial Life)*, ser. LNAI, J. Kelemen and P. Sosik, Eds., vol. 2159. Springer, 2001, pp. 704–713.

[25] D. Polani, C. Nehaniv, T. Martinetz, and J. T. Kim, "Relevant information in optimized persistence vs. progeny strategies," in *Proc. Artificial Life X*, L. M. Rocha, M. Bedau, D. Floreano, R. Goldstone, A. Vespignani, and L. Yaeger, Eds., August 2006, pp. 337–343.

[26] M. Saerens, Y. Achbany, F. Fuss, and L. Yen, "Randomized shortest-path problems: Two related models," *Neural Computation*, vol. 21, pp. 2363–2404, 2009.

[27] N. Tishby and D. Polani, "Information theory of decisions and actions," in *Perception-Action Cycle: Models, Architecture and Hardware*, V. Cutsuridis, A. Hussain, and J. Taylor, Eds. Springer, 2010, in Press.

[28] C. Paul, "Morphology and computation," in *Proceedings of the International Conference on the Simulation of Adaptive Behavior, Los Angeles, CA, USA, July*, 2004.

[29] ——, "Morphological computation: A basis for the analysis of morphology and control requirements," *Robotics and Autonomous Systems*, vol. 54, no. 8, pp. 619–630, 2006.

[30] R. Pfeifer and J. Bongard, *How the Body Shapes the Way We think: A New View of Intelligence*. Bradford Books, 2007.

[31] R. S. Sutton and A. G. Barto, *Reinforcement Learning*. Cambridge, Mass.: MIT Press, 1998.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[33] L. M. Rocha, M. Bedau, D. Floreano, R. Goldstone, A. Vespignani, and L. Yaeger, Eds., *Proc. Artificial Life X*, August 2006.