van Dijk, S., and Polani, D., (2011). Look-Ahead Relevant Information: Reducing Cognitive Burden over Prolonged Tasks. In *Proc. IEEE Symposium Series in Computational Intelligence 2011 — Symposium on Artificial Life*, 46–53. IEEE.

# Look-Ahead Relevant Information: Reducing Cognitive Burden over Prolonged Tasks

Sander G. van Dijk
Daniel Polani
Adaptive Systems Research Group
Univeristy of Hertfordshire
Hatfield, UK
Email: s.vandijk@herts.ac.uk

*Abstract*—**Based on the fact that information processing is costly, we study in this paper the trade-off between performance and informational requirements. Most importantly, we are interested in how local decisions can alleviate future cognitive burden, measured by the amount of sensory information an agent processes, without conceding performance. We introduce *look-ahead information* as a novel concept to capture the long-term informational requirements and present an iterative method to determine the value of this quantity. Using an example problem, we show how these long-term considerations enable an agent to predict future effects of its actions on its informational burden, and to shape the course of the world to achieve more informationally parsimonious behaviour.**

## I. Introduction

Spending energy is costly for living organisms. Any extra energy spent on detrimental, or even on simply non-beneficial behaviour may put the organism at a disadvantage compared to its competitors in the struggle for life. Even if the behaviour performed by the organism is desirable, it is still disadvantageous to spend more energy on this behaviour than the minimum that an agent can get away with. So not surprisingly, examples of organisms that behave in the most efficient way are abundant, from load-carrying by starlings and honey bees to prey size selection by crabs [1].

However, it does not only pay to be physically conservative, a lot of energy can also be saved by being mentally parsimonious. It is well known that information acquisition and processing are an expensive part of an organism's life. From blow fly to human beings, it has been shown that a significant part of the energy consumption of an organism is accounted for by its sensors and/or information processing organs, i.e. brain [2], [3].

Because of this high cost, it is plausible that an organism will attempt to keep the informational requirements of its behaviour as low as possible; when confronted with a task, it may aim to perform this task not only physically but also mentally in the least demanding way. For example, during navigation, an organism could choose to take a route to its destination that may take longer than an alternative path, but which involves much less crucial navigation points that would require a much higher level of attention, such as having many turns, or areas where it is easy to make a detrimental step if the organism is distracted and does not have taken in enough information.

It is becoming increasingly popular to study the informational properties and structures underlying the behaviour and morphology of agents, however this is certainly not a new approach. Already 50 years ago for instance, Barlow hypothesised that the function of sensory relays can be explained as informational restructuring and redundancy reduction [4]. More recently, informational approaches have been used to study and explain, amongst other things, adaptive rescaling of sensory systems [5], sensory ecologies [6] and embodiment [7]. Some behaviour found in nature can even be fully understood in terms of information [8].

The formal foundation of these results is provided by the field of information theory, which was initiated by Shannon [9]. Although the theory explicitly ignores the semantics and value of information in its initial formulation, the work referenced in the previous paragraph shows that it is certainly possible to use information theoretical concepts in (artificially) lifelike settings, where meaning and relevance is important. A prime example of this is the notion of *relevant information* [10], the main basis of the work presented in this paper. As we will show in more detail below, this concept enables one to exactly quantify the amount of information an agent on average needs to process in order to perform a certain task to a fixed level of performance, which is a fundamental invariant of a task and the environment in which it is to be performed, and to determine a policy that achieves this minimum.

The concept of relevant information is based on the hypothesis alluded to in the beginning of this section: that agents aim to minimize information intake needed to achieve a certain level of performance, due to the high energy costs of information processing. This concept already offers powerful tools to analyse and explain behaviour and can be applied to different sources of information, e.g. sensoric information [10], or information about an agent's goal [11]. However, the original formulation of relevant information only considers short time scales and is inherently reactive; when searching for a policy that minimizes the average amount of information taken in, it does not take into account that local changes in behaviour can greatly affect the future course of the world. Therefore, it fails to find policies that may require some decisions that are *locally* more costly, but that lead the agent through a part of the environment that is much less informationally demanding *in the long run*.

Here we will address this issue by extending the concept of relevant information and introducing the novel informational quantity of *look-ahead relevant information*, develop and discuss methods to find a policy that achieves the minimum of this new quantity, and show how minimization of this quantity affects the overall average informational burden of an agent *quantitatively*, and the structure of its behaviour *qualitatively*.

## II. Perception-Action Loop

As mentioned in the introduction, the relevant information is the minimum amount of information that an agent needs to take in and process to achieve a certain level of performance. This abstract description signals that one must be able to quantify two things to derive a concrete concept: *information intake* and *performance*.

To enable us to do this, we firstly set up a more formal model of the interactions between an agent and the environment. An agent is equipped with sensors and actuators that connect it to the world. At any given time, the agent senses the state of the world $W$, acquiring a sensory state $S$. Based on this sensation the agent selects and performs an action $A$. This action influences the state of the world, which in its turn causes a new sensory state in the agent. This loop, shown in Fig. 1(a), continues until the agent dies, and is called the *Perception-Action loop* (PA-loop). Note that here we assume that the agent is purely reactive; it has no memory and its actions are completely determined by the current sensory state.

If we assume that the world is fully accessible to the agent, i.e. it can sense the full state of the world, the loop can be simplified by collapsing the world-state and sensory-state nodes. If we then unroll the PA-loop in time we arrive at the acyclic graph of Fig. 1(b). By treating the nodes as random variables, the PA-loop is now modeled as a *Causal Bayesian Network* (CBN) [12].

The edges of the CBN show the causal relations between the different variables in the PA-loop: the state of the world determines which action the agent takes, and the next state of the world is determined by the execution of this action and the previous state. In a CBN, these causal interactions are not limited to being deterministic. The agent does not have to select the same action every time for a certain state, and performing an action in a certain state may have different, e.g. randomized, outcomes. Such stochastic interactions are described by probability distributions. Firstly, the probability of selecting action $a_t$ in state $s_t$ is given by the conditional probability distribution $\pi(a_t|s_t)$, which is called the agent's *policy*. Secondly, $p(s_{t+1}|s_t, a_t)$ denotes the probability of the world transiting to the new state $s_{t+1}$ when the agent performs action $a_t$ in state $s_t$. The combined dynamics of the policy and the state transition probability distribution determine the development of the state of world, and thus the probability $p(s_t)$ of the agent arriving at a certain state.

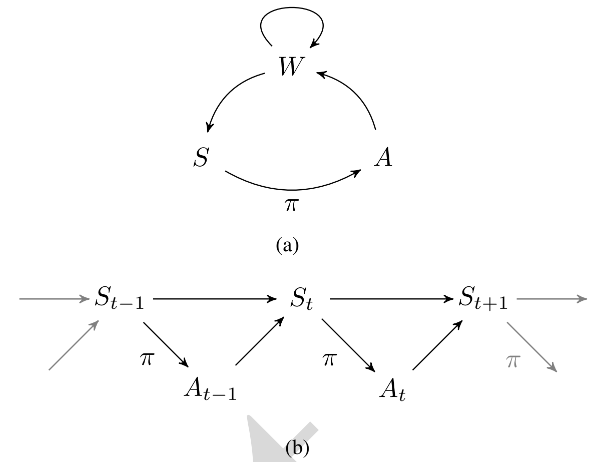


Fig. 1. (a) Perception-action loop, and (b) Causal Bayesian network of the perception-action loop with a fully accessible world, unrolled in time. The state of the world at time $t$ is denoted by $S_t$, the action selected by the agent, according by its policy $\pi$, by $A_t$.

## III. Information Intake

Before knowing in which particular state the environment currently is, an agent will generally have a high uncertainty about what action it should perform. The best strategy would be to select an action based on the distribution of actions over all states; e.g. if its policy dictates a certain action in 4 out of 5 states, a blindfolded agent would select this action with probability 0.8. Now assume that the agent is allowed to sense the current state after all. With this it has more information on which to base its action selection, and the uncertainty about what to do decreases. This drop in uncertainty can be used to measure how much information the agent actually acquired and needed to process to be able to select the correct action. It is this amount that can be correlated to the cognitive burden of the agent [10], [13], and as we will show here, the metaphor of information channels enables us to apply concepts from the field of information theory to study it, and to quantify it exactly.

The average probability of taking a certain action is given by the a-priori distribution $p(a_t) = \sum_{s_t} p(s_t)\pi(a_t|s_t)$. Using this distribution, the pre-sensing action uncertainty is measured by the *entropy* $H(A_t) = -\sum_{a_t} p(a_t) \log p(a_t)$. If the logarithm is taken with base 2, which we will do in the remainder of this paper, entropy, and quantities derived of entropy, are measured in *bits*. After sensing $s_t$, the agent now knows the actual appropriate distribution over actions determined by its policy $\pi(a_t|s_t)$. The new uncertainty is measured by the conditional entropy $H(A_t|s_t) = -\sum_{a_t} p(a_t|s_t) \log p(a_t|s_t)$. The drop in uncertainty, which as discussed above is equivalent to the amount of information taken in and processed by the agent to decide its action, is then the difference between the a-priori and conditional entropies: $I(s_t; A_t) = H(A_t) - H(A_t|s_t)$. Finally, the *average per-step information intake* is the average of this quantity over all states, weighed by the probability of the agent arriving in these states:

$$I(S_t; A_t) = \sum_{s_t} p(s_t) I(s_t; A_t) \tag{1}$$

This quantity is called the *mutual information* between $S_t$ and $A_t$, is symmetric and always non-negative [14].

When we look at extreme cases, it is intuitive to see that this quantity is indeed correlated with the perceived cognitive complexity of the policy that determines its value. Firstly, we consider the case when $H(A_t)$ equals zero, so when there is already no uncertainty about which action to take before having sensed anything. Since uncertainty cannot be negative, no extra information can decrease it any further, and the agent can act blindly with zero information intake. This is indeed indicated by the fact that here $I(S_t; A_t)$ equals 0, due to the non-negativity of mutual information. The only policies that achieve this extreme case have a very simple form: they dictate the same single action for each state.

Next, we consider the case where $H(A_t)$ is non-zero, and can even be very high, but where $H(A_t|s_t)$ is such that $\sum_{s_t} p(s_t)H(A_t|s_t) = H(A_t)$. This means that sensing the state of the world does on average not decrease the uncertainty about what action to take, and that here, too, $I(S_t; A_t)$ equals 0. If we look at which policies actually achieve this, we see again that they are intuitively simple: the agent can now choose from a larger set of actions, but the distribution over these actions, i.e. the policy, is the same for each state.

So, in the previous two cases the agent does not have to discern which state the agent actually is in, because what to do is the same in each state anyway. This changes in the last extreme case, where $H(A_t)$ is maximal and $\sum_{s_t} p(s_t)H(A_t|s_t) = 0$. Here, the decrease in uncertainty caused by sensing the state of the world is large and $I(S_t; A_t)$ is at its maximum. The policy that achieves this puts a high cognitive burden on the agent: it selects a single unique action for each state, so that the agent has to take in all information that is available to decide exactly which state the world is in, to make sure it performs the correct action.

Thus, we can conclude that minimizing (1) leads to intuitively simpler policies that alleviate the agent's cognitive burden by requiring less intake and processing of information.

## IV. Relevant information I: Overview

Just directly minimizing this quantity, however, will generally not result in a successful agent. As mentioned earlier, often the performance of an agent needs to be up to a certain level. For instance, when the agent attempts to navigate to a certain destination, it will commonly strive to get there as fast as possible. Performance is then measured by the time it takes the agent to achieve its desired state. To ensure this level of performance the agent will generally not be able to decrease its information intake to zero. However, certainly not all information that is available to an agent is actually relevant to its performance, and can be ignored to save mental processing capacity. The colour of the trees in a forest for instance is not relevant to an organism navigating through it, it only needs to sense and process *where* the trees actually are.

Let us now assume that from all the policies an agent could follow, there is a small set that achieve a certain desired level of performance. The amount of *relevant information* is then defined as the minimum amount of information that is processed by the agent, over this set of policies [10]:

$$I(S; A^*) = \min_{\pi:\pi \text{ achieves desired performance}} I(S; A). \tag{2}$$

This quantity is a fundamental invariant of the agent-environment dynamics and the task that the agent attempts to perform; if the agent is not able to, or just does not, take in and process this amount of information, e.g. when its sensors do not have high enough resolution, or the bandwidth of its information processing system is too low, the agent will in no way be able to achieve the desired performance.

## V. Performance

Now we come to the second problem: how to actually quantify performance. To do this, we can draw from the field of Reinforcement Learning (RL), and define a reward function $R^{a_t}_{s_t,s_{t+1}}$ [15]. This function determines a reward $r_t$ given to the agent when it performs action $a_t$ in state $s_t$, and by doing so arrives in state $s_{t+1}$. Using this, we can define the *utility* of performing an action in a given state as the total expected future reward obtained by the agent by doing so, and consecutively following a certain policy $\pi$:

$$\begin{aligned} U^\pi(s_t, a_t) &= E_\pi\Big[\sum_{k=t}^{\infty} r_k\Big] \\ &= \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t) \cdot \\ &\qquad \Big[R^{a_t}_{s_t,s_{t+1}} + E_\pi[U^\pi(s_{t+1}, A_{t+1})]\Big] \end{aligned} \tag{3}$$

The recursive, so called *Bellman* form of $U^\pi(s_t, a_t)$, where the value is determined by the sum of an immediate element and a future expectation of the function, is common in the field of reinforcement learning [15], and we will encounter it again in section VII. With (3) established, we can now measure performance as the expected utility achieved by an agent's policy:

$$E_\pi[U^\pi(S_t, A_t)] = \sum_{s_t,a_t} p(s_t)\pi(a_t|s_t)U^\pi(s_t, a_t). \tag{4}$$

## VI. Relevant Information II: Computational Details

Now that we are able to quantify both information intake and performance, we can finally determine the amount of relevant information for a certain level of performance [10]. The 'desired performance' of problem (2) can be equated to a specific value of the expected utility, resulting in a concrete constraint on the policy. The problem now becomes the problem of minimizing the average information intake $I(S_t; A_t)$ while fixing the expected utility $E[U^\pi(S_t, A_t)]$.

With the method of Lagrange multipliers this problem can be turned into an unconstrained minimization problem:

$$\min_{\pi}\Big[I(S_t;A_t)-\beta E[U^{\pi}(S_t,A_t)]\Big]. \tag{5}$$

The Lagrange multiplier $\beta$ implicitly encodes the constraint on performance. For high values of $\beta$, utility becomes more important, and in the limit $\beta \to \infty$ the possible policies are limited to ones that achieve the highest performance possible, or *optimal policies*. When $\beta$ is fixed to lower values, however, the focus is instead on information parsimony.

Problem (5) has a similar form to the classical rate-distortion problem, well known in the field of information theory. This problem consists of finding a channel with the smallest possible bandwidth that does not cause more than a desired amount of distortion, i.e. wrong values in the output. A solution to this problem can be found with the Blahut-Arimoto algorithm [16]. In our case, the channel is formed by the agent's policy, the bandwidth is the amount of sensory information that is taken in and processed to achieve a particular utility, and distortion is equated to the negative of the expected utility, which quantifies the average 'wrongness' of an action. However, in the rate-distortion problem distortion measures are fixed, whereas the utility is dependant on the policy and needs to be kept consistent with the policy during optimization to achieve sensible results. This is done by interleaving the policy iterations from the Blahut-Arimoto algorithm derived from (5) with value updates according to (3), resulting in the following process:

$$\to \pi \xrightarrow{(3)} U^{\pi} \xrightarrow{(5)} \pi' \to . \tag{6}$$

This process is iterated until convergence of the policy.

## VII. Look-Ahead Relevant Information

The relevant information found with this method depends on the final policy. However, both $I(S_t;A_t)$ and $E[U^{\pi}(S_t,A_t)]$ are not just determined by which actions the agent takes in certain states, but also by the states it is actually likely to find itself in. It could be that the utility of actions in a certain state are very low, but if a policy is chosen such that the agent never visits this state, this utility does not contribute to the overall expectation. As an example, consider an organism crossing a river at a ford to avoid having to spend a lot of energy on swimming across. Similarly, the agent could possibly decrease its overall average informational burden by avoiding more complex states, e.g. by following a well-trodden, easy to recognize path instead of navigating through a dense forest.

More concretely, the states in which an agent is likely to find itself depends on the state distribution $p(s_t)$. The traditional concept of relevant information assumes that this distribution is fixed, and usually defines it to be a uniform distribution. This aims to ensure the usability of the Blahut-Arimoto update steps as much as possible, by staying close to the form of the rate-distortion problem for which these steps were designed. However, such a distribution will in general not be consistent with the agent's policy and reflect the true effect of the policy on the course of the world. It is likely that this limitation results in suboptimal information-performance trade-offs. As we will see later, a simple improvement that gives more sensible results is to compute a state distribution that is consistent to the current policy at each iteration of 6, just as how the utility function is kept consistent.

However, this does not take away the more fundamental limitation of relevant information, that it does not take into account that an agent will generally be able to shape the distribution over states it will visit in order to increase its expected utility and/or decrease its average cognitive burden. The relevant information framework as described in the previous section only considers short term, *single-step* dynamics.

It is likely then, that by taking into account the effects that changes in a policy have on the state distribution, an agent can settle on policies that require less information on average to achieve the same performance, or, equivalently, policies that achieve higher performance with the same average information intake. However, because of the intricate causal connections in the PA-loop, it is very difficult to determine the exact effects of actions on the future of the world, and solving (5) directly quickly becomes infeasible in larger environments. Therefore, we here introduce a heuristic method to estimate the solution.

For the effect of actions on the future performance of the agent there is already a heuristic in place. In each iteration of 6, the new policy $\pi'$ is chosen to get closer to the desired expected future sum of rewards. However, for this sum the utility of actions under the assumption of the old policy $\pi$ is used as a heuristic. We introduce the *look-ahead information* $\mathfrak{I}^{\pi}(s_t)$ as a similar heuristic for the effect of policy changes on the average amount of information intake.

The look-ahead information defines the informational cost of being in a state $s_t$ as the sum of the immediate contribution to the total information intake, $I(A;s) = H(A) - H(A|s)$, and the expected accumulated intake over the remainder of the run. This allows us to write the quantity as a recursive equation in Bellman form, in a way similar to the formulation of the utility function in (3):

$$\mathfrak{I}^{\pi}(s_t) := H(A_t) - H(A_t|s_t) + E_{\pi}[\mathfrak{I}^{\pi}(S_{t+1})]. \tag{7}$$

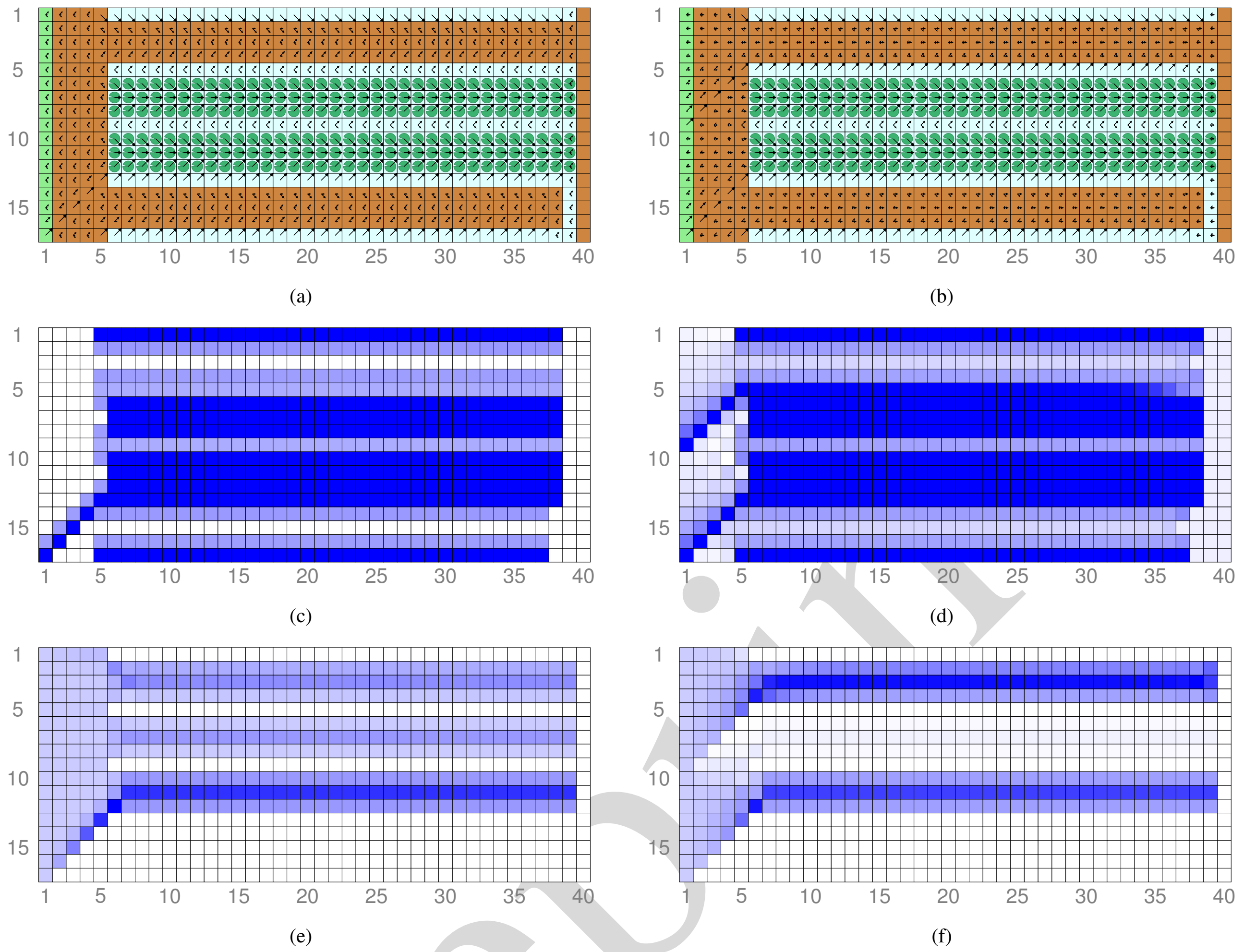
The formulation of informational terms in this form has recently been developed to describe the total information gain of an agent-environment combination, termed the information-to-go [13]. Here we use this form in a novel, agent-centred way to quantify its long-term informational burden.

Analogous to the single-step case, we define the *look-ahead relevant information* as the minimum of the look-ahead information over the set of policies that achieve a certain desired level of performance:

$$\mathfrak{I}^{*}(s_t) = \min_{\pi:\pi \text{ achieves desired performance}} \mathfrak{I}^{\pi}(s_t), \tag{8}$$

and, via Lagrange, derive the new unconstrained formulation:

$$\min_{\pi}\Big[E_{\pi}[\mathfrak{I}^{\pi}(S_t)] - \beta E_{\pi}[U^{\pi}(S_t,A_t)]\Big]. \tag{9}$$



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 2. The navigation scenario used in the experiments. An agent is placed on one of the west-most cells and has to cross over to the other side of the world, onto one of the east-most cells. Pieces of land are marked with a brown background, and water is denoted by blue. The water cells in rows 6-8 and 10-12 are covered by green lily pads. The agent can move in three ways: jump north-east, east, or south-east. Jumps from land are deterministic, they result in the agent moving in the intended direction. Jumps made from water are more noisy, 50% of the time these result in the agent landing one cell north or south of where it would normally land. Every jump costs the agent 1 point, however jumping from the water is more difficult, so the agent endures a cost of 10 points when landing into it. The left three panels show the results obtained with the traditional relevant information methods, the others the results obtained using look-ahead relevant information. Panels (a) and (b) show the optimal policies found with these methods (i.e. in the limit $\beta \to \infty$), where the length of the arrows in a cell correspond to the probability of taking the action that aims to move the agent into that direction in that state. The local information intake $I(s_t; A_t)$ that result from these policies is shown for each state in panels (c) and (d), with higher values being darker. The last two panels show the relative probability $p(s_t)$ for each state that the agent will visit that state in a run using these policies. Higher probabilities are denoted with darker shading.

The algorithm for the single-step relevant information can be easily extended to the look-ahead case by replacing the single-step cost $H(A_t) - H(A_t|s_t)$ by $\mathfrak{I}^\pi(s_t)$ in the Blahut-Arimoto policy update steps in (6), and by updating $\mathfrak{I}^\pi(s_t)$ at each iteration according to (7). Thus, we perform the following process until convergence:

$$\to \pi \xrightarrow{(3),(7)} U^\pi, \mathfrak{I}^\pi \xrightarrow{(9)} \pi' \to . \qquad (10)$$

## VIII. Experiments

To display the effects of a drive to minimize the look-ahead information to the relevant minimum, we present an example scenario in the form of a navigation problem. In this scenario, the agent has to traverse the environment, which offers several pathways that the agent can take. Some of these pathways are optimal, whereas one results in the agent enduring a higher cost. Also, some paths are deterministic, while other paths are more noisy and require more information to perform the optimal policy. These difference ensure that the agent can increase performance, decrease informational requirements, or settle on a trade-off by changing the distribution of the paths it will take to get to the final goal.

More concretely, the agent is placed in a world that consists of a rectangular grid of 40×17 cells, as shown in Fig. 2. The world contains a patch of land 5 cells wide in the west and a line of land in the east. There are two land pathways between these pieces of land, in rows 2-4 and 14-16. The southern of the two is cut off by a line of water at the end. Two additional pathways are formed by lily pads, covering rows 6-8 and 10-12. The four pathways are divided by three lines of open water.

van Dijk, S., and Polani, D., (2011). Look-Ahead Relevant Information: Reducing Cognitive Burden over Prolonged Tasks. In *Proc. IEEE Symposium Series in Computational Intelligence 2011 — Symposium on Artificial Life*, 46–53. IEEE.

The state of the world at a given time, $S_t$, is the location of the agent in the world, which can be in any of the 680 cells in the grid.

In a single run, the agent is placed in one of the west-most cells (column 1, marked green). The goal of the agent is to reach the shore in the east (column 40). At each time step the agent can select one of 3 actions: jump either one cell north-east, one cell east, or one cell south-east. The world is fully surrounded by walls; performing an action that would have the agent run into the wall results in the agent moving simply eastwards.

Each jump consumes energy, incurring a cost to the agent. This cost is represented by a negative reward: $R^{a_t}_{s_t,s_{t+1}} := -1$. Moving is more costly when the agent has fallen into the water. In this case, the cost goes up to 10, i.e. here $R^{a_t}_{s_t,s_{t+1}} := -10$. The agent can prevent this by hopping onto the lily pads that float on the water. Finally, the reward is 0 when the agent arrives in one of the goal states, to mark that it has finished the task and to limit the total cost.

A policy thus is optimal when it brings the agent to the other side, without falling into the water. This is achieved by following the northern land path, or one of the two paths formed by lily pads. However, the lily pads are unstable, making the effect of a jump from one uncertain. With a probability of 0.5, such a jump results in the agent landing either one cell further north or one cell further south than where the action would normally take the agent. The same indeterminacy holds when the agent attempts to jump from open.

This means that on the two pathways formed by lily pads the agent has to be extra careful not to end up in the water. In fact, on these pathways there always is only a single optimal action available: when next to the open water, try to move away from it, otherwise try to move straight ahead. Any other strategy has the risk of diverting the agent into the water. This means that it has to pay close attention to where it is, to be able to select the correct action.

On the two outer paths, however, the structure of the world offers the agent help to alleviate its cognitive burden. Here, the lack of noise allows it to venture closer to the water and to worry less about which action to take. In each cell on these pathways there are multiple actions that ensure that the agent will not get wet.

In this environment we perform three experiments. Firstly, we determine a policy following the original single-step relevant information method, for different $\beta$ values, using a uniform state distribution for each iteration of (6). This is the method originally introduced by Polani et al [10]. Secondly, we will perform the same experiment, but with the added step of making the state distribution consistent to the current policy at each iteration. To differentiate these experiments, we will refer to the first as the *inconsistent* single-step case, due to the fact that the uniform state distribution that is used generally is not consistent with the policy that is considered. Finally, 1 the experiment is repeated using look-ahead information and consistent state-distributions.

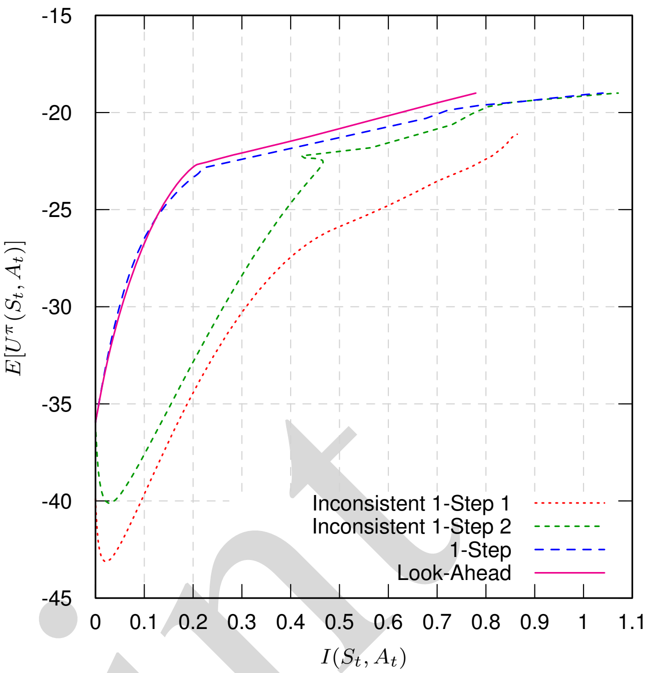


Fig. 3. Trade-off curves for single-step and look-ahead relevant information in the environment of Fig. 2. These curves show the information-utility trade-off found for different values of $\beta$, for the single-step and look-ahead relevant information methods, and therefore the maximum average utility obtainable when the average per-step information intake is limited to a certain amount, or the minimum average information intake needed to achieve a certain utility level. The end-points of the 1-step and look-ahead curves correspond to the policies shown in Figs. 2(a) and 2(b).

## IX. Results

In this section we will present and compare the results of the three experiments described above. For the final policies that are found we determine the average per-step information intake, $I(S_t; A_t)$, and the performance, measured by the expected utility $E_\pi[U^\pi(S_t, A_t)]$. Doing so for a range of values of $\beta$ results in an information-performance trade-off curve.

There are certain properties that such a curve should have to be correct. Most importantly, it must be monotonic, because each policy that is feasible with a strong restriction on information intake, and with it its corresponding performance, is also available when this restriction is weakened; allowing more information intake can only expand the set of feasible policies and thus cannot decrease the maximum possible performance.

When we examine the trade-off curves resulting from the three different experiments, which are shown in Fig. 3, we find that the curve obtained with original relevant information does not have this shape. This curve, marked 'Inconsistent 1-Step 1', shows a dip at low values of information intake in our scenario, suggesting that here more information causes the agent to perform worse. As deduced earlier, the policies that correspond to the dip in the trade-off curve can therefore not be optimal. Indeed, analysis of these solutions indicates that they do not achieve a (local) minimum of the Lagrangian function of (5); Monte Carlo sampling near the final policies yields both solutions that give higher as solutions that give lower values for this function, suggesting that here the algorithm converged to a saddle point. This also holds for solutions further along the curve, approximately up to where the average information intake surpasses 0.45 bit. The same analysis indicates that the policies corresponding to larger amounts of information intake, and those found with the lowest $\beta$ values, *are* optimal, at least locally.

This shows that using the original single-step relevant information methods does not always find a correct trade-off curve. Another limitation of these methods is that utility and information intake values that make up the the curve are calculated using the same uniform state distribution as used during the iteration of (6). These values often do not reflect the actual ones, since the uniform distribution is generally not consistent with the final policy. The actual trade-offs, calculated using the state distributions that are consistent with the policies, are shown by the curve marked ‘Inconsistent 1-Step 2’ in Fig. 3. This curve shows that the actual optimal performance level in this scenario is higher than estimated by the original method, -19 instead of -21.12, but also that the actual information intake needed to achieve this optimum is higher, 1.07 instead of 0.86 bit. Furthermore, the section where the method gets stuck on saddle points is more clearly visible, the border of which is marked by the incoherent artefact between the 0.4 and 0.5 bit marks.

Next, the curve marked ‘1-Step’ shows the trade-off found in the second experiment, where a consistent state distribution is not only used for determining the final trade-off, but also for each iteration of the single-step relevant information method. One sees that this small improvement restores the expected monotonicity property of the curve, and causes the algorithm to generally converge to better trade-offs; policies are found that require less information to achieve the same level of performance. The optimal policy found in this case, corresponding to $\beta$ approaching infinity, has a decrease in information intake of just 0.03 bit compared to the inconsistent single-step case, and when $\beta$ approaches 0 both methods converge to the same policy, but for intermediate values for $\beta$ the difference becomes more significant.

Finally, the trade-off curve found with look-ahead information is marked with ‘Look-Ahead’ in Fig. 3. These results show that the ability to estimate the effect of actions on future informational requirements enables an agent to decrease the average per step information, while maintaining the same performance level. The optimal policy found in this experiment achieves an average information intake of 0.78 bit, a drop of 25% compared to the optimal policy found with the single-step method.

These two optimal policies are shown in Figs. 2(a) and 2(b), and as expected share properties that make them optimal. Firstly, we observe that in both cases the optimal policy for the middle two pathways, created by the lily pads, is as was discussed in the previous section: the agent moves away from the water cells if it is next to it, otherwise it moves straight on. Secondly, also as expected, both cases show a more stochastic policy on the land pathways. As can be seen in Figs. 2(c) and 2(d), where the local information intake $I(s_t; A_t)$ is shown, this difference in the policies for the land and lily pad pathways cause a large difference in the informational burden; the instability of the lily pads cause a higher amount of required information intake. Secondly, in both cases the agent shows a clear preference for the southern lily pad pathway over the suboptimal southern land route, since the latter forces the agent to go through one cell of water at the end.

The main qualitative difference in the two policies, which results in the quantitative difference in informational burden, is visible in the northern part of the environment. An agent concerned only with single-step relevant information has the same probability of taking the northern land pathway as of taking the northern lily pad pathway, whereas an agent that looks ahead takes the land pathway whenever possible. This results in a large shift in the state distribution, shown in Figs. 2(e) and 2(f), and because of the large difference in information requirements along the two paths, as shown in Figs. 2(c) and 2(d), in the significant drop in informational requirements.

When the requirement of optimal performance is lifted by decreasing $\beta$, the relative information drop between the single-step and look-ahead cases slowly goes down, to approximately 10% when performance has decreased to -22.8, and the difference disappears at an expected utility of -25. When analysing the policies and state distributions along these segments of the curves, we find that the advantage of looking ahead is maintained here by having a preference for traversing the land pathways through the centre (rows 3 and 15), while in the single-step case the agent moves more often along the water. This is beneficial because at the middle of these pathways any action is optimal, so the agent can choose any policy here that requires little information. This effect can be seen clearly in Fig. 2(c), which shows that the information intake in the middle of the land-paths is lower than at the outsides.

Finally, for the lowest values of $\beta$ the single-step curve lies slightly left of the look-ahead curve; here the estimation of future informational costs actually result in policies that require a fraction of information more to achieve the same performance.

## X. Discussion

At first glance, the absolute gain in information parsimony of around 0.26 for optimal policies does not seem large. However, this is partly due to the relative simplicity of the scenario: at any state no more than 3 actions are optimal, putting a strict upper bound of 1.58 bit on the amount of relevant information. When taking this into account, a drop of 0.26 bit corresponds to 16% of the maximum. In more complex scenarios, most notably ones with a larger action space, the absolute gains are expected to be more significant. Secondly, we find that a future decrease of required information is paid for by increasing the complexity of the action selection in earlier stages. In the scenario presented in this paper for instance, the agent needs to process more state information in the first five steps to make sure it ends up in the pathway that presents the lowest long-term burden, as shown in Figs. 2(c) and 2(d).

This puts a limit to the advantage that can be obtained by looking ahead, and the difference in trade-offs found for lower values of $\beta$ lead to another important conclusion: although a significant gain can be achieved by considering future effects when optimality is important, an agent could already do well with simpler short-term strategies when performance close to the optimum is not required. In this case, one can argue that the decrease in required information acquisition and processing capabilities is outweighed by the extra cognitive burden resulting from having to maintain estimations of long-term effects. As discussed in the last paragraph of the previous section, the results offered in this paper show that these estimates can even cause an agent with a high constraint on information intake bandwidth, i.e. for the lowest values for $\beta$, to settle for a slightly lower performance than actually possible.

This effect may be caused by the algorithm converging to local minima of (9), or it could be a result of the look-ahead information as currently formulated being an incomplete heuristic for the actual effects of local actions on global information intake. Additional research is required to study this effect further, and, more generally, to be able to determine whether the trade-offs found by any method are fully optimal. Although numerical analysis, through Monte Carlo sampling, of the solutions found for the single-step and look-ahead case indicate that these are at least local minima of the problems they solve, respectively (5) and (9), the results obtained in the previous section show that neither formulation of the problem captures the problem of determining the true optimal trade-off in all cases.

## XI. Conclusion

In this paper, we have discussed the limitations of the original formulation of relevant information. We have shown that this can lead to the failure of recovering the optimal trade-off between cognitive burden, measured by the amount of information required to perform a task, and the achieved level of performance on that task. To be able to improve this trade-off, we have introduced the novel concept of look-ahead relevant information. We have given methods to make the computation of this quantity feasible, and to find policies that achieve this minimum amount of information intake. This enables one for the first time, to the knowledge of the authors, to study the effect of the estimation of long-term informational effects of actions on the trade-off between performance and information processing demands. We have shown that the policies found by this new method can achieve a decrease of cognitive burden without conceding performance by making the agent avoid states that are informationally costly, resulting in informationally more parsimonious behaviour.

## References

[1] J. R. Krebs and N. B. Davies, *An introduction to behavioural ecology*, 3rd ed. Blackwell Scientific Publications, Oxford, England, 1993.

[2] S. B. Laughlin, R. R. de Ruyter van Steveninck, and J. C. Anderson, "The Metabolic Cost of Neural Information," *Nature Neuroscience*, vol. 1, pp. 36–41, 1998.

[3] D. Polani, "Information: Currency of Life?" *HFSP Journal*, vol. 3, pp. 307–316, 2009.

[4] H. B. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," in *Sensory Communication*, W. Rosenblith, Ed. Cambridge, MA: MIT Press, 1961, ch. 13, pp. 217–234.

[5] N. Brenner, W. Bialek, and R. de Ruyter van Steveninck, "Adaptive Rescaling Maximizes Information Transmission," *Neuron*, vol. 26, no. 3, pp. 695–702, 2000.

[6] C. L. Nehaniv, D. Polani, L. Olsson, and A. S. Klyubin, "Information-Theoretic Modeling of Sensory Ecology: Channels of Organism-Specific Meaningful Information," in *Modeling Biology: Structures, Behaviors, Evolution (The Vienna Series in Theoretical Biology)*, M. D. Laubichler and G. B. Müller, Eds. MIT Press, 2007, pp. 241–282.

[7] R. Pfeifer, M. Lungarella, O. Sporns, and Y. Kuniyoshi, "On the Information-Theoretic Implications of Embodiment – Principles and Methods," in *Proc. of the 50th Anniversary Summit of Artificial Intelligence*, vol. 4850. Springer-Verlag, 2007, pp. 76–86.

[8] M. Vergassola, E. Villermaux, and B. I. Shraiman, "'Infotaxis' as a Strategy for Searching Without Gradients." *Nature*, vol. 445, no. 7126, pp. 406–9, 2007.

[9] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, 1948.

[10] D. Polani, C. L. Nehaniv, T. Martinetz, and J. T. Kim, "Relevant Information in Optimized Persistence vs. Progeny Strategies," in *Artificial Life X: Proceedings of The 10th International Conference on the Simulation and Synthesis of Living Systems, Bloomington IN*, 2006.

[11] S. G. van Dijk, D. Polani, and C. L. Nehaniv, "What do You Want to do Today? Relevant-Information Bookkeeping in Goal-Oriented Behaviour," in *Artificial Life XII: The 12th International Conference on the Synthesis and Simulation of Living Systems*, H. Fellermann, M. Dörr, M. Hanczyc, L. L. Ladegaard, S. Maurer, D. Merkle, P.-A. Monnard, K. Støy, and S. Rasmussen, Eds. Odense, Denmark: The MIT Press, Cambridge, Massachusetts, 2010, pp. 176–183.

[12] 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*, R. Zebulum, D. Gwaltney, G. Hornby, D. Keymeulen, J. Lohn, and A. Stoica, Eds. Los Alamitos, CA: IEEE Computer Society, 2004, pp. 177–180.

[13] N. Tishby and D. Polani, "Information Theory of Decisions and Actions," in *Perception-Reason-Action Cycle: Models, Algorithms and Systems*, V. Cutsuridis, A. Hussain, and J. Taylor, Eds. Springer (In Press), 2010.

[14] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[16] R. E. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Transactions on Information Theory*, vol. 18, pp. 460–473, 1972.