What do You Want to do Today?
Relevant-Information Bookkeeping in Goal-Oriented Behaviour

Sander G. van Dijk, Daniel Polani and Chrystopher L. Nehaniv

Department of Computer Science, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
s.vandijk@herts.ac.uk

Abstract

We extend existing models and methods for the informational treatment of the perception-action loop to the case of goal-oriented behaviour and introduce the notion of *relevant goal information* as the amount of information an agent necessarily has to maintain about its goal. Starting from the hypothesis that organisms use information economically, we study the structure of this information and how goal-information parsimony can guide behaviour. It is shown how these methods lead to a general definition and quantification of sub-goals and how the biologically motivated hypothesis of information parsimony gives rise to the emergence of properties such as least-commitment and goal-concealing.

Introduction

The world is a complex place. Millions of years of evolution have created an environment with intricate relationships, structure and many things that an organism living in it has to look out for. It is no surprise then that organisms invest a lot of energy in the processing of all the information available to them. For instance, the retina of a resting blowfly accounts for 10% of its energy consumption and for the human brain this amount is estimated to be 20% (Laughlin et al., 1998).

It is unlikely that an organism would spend all this energy if it is not important; individuals that limit their information intake and processing to the necessary minimum and allocate the rest of their energy to behaviour that is more relevant to survival or reproduction will outperform ones that waste energy on useless information processing. Also, even though this means an organism uses information economically, it is plausible that an organism still often operates at the limit of its information processing bandwidth and that there is an evolutionary drive to do away with unused capacity, similar to the degeneration of useless eyes in cave-dwelling fish (Jeffery, 2001). We will refer to these assumptions as the *information parsimony* hypothesis.

We are interested in the necessary principles of life and lifelike behaviour. The hypothesis of information parsimony hints that information acquisition and processing capabilities are part of these fundamental requirements. In the vein of the Alife motto "life as it could be", we use minimal models of agents and their informational properties to study these basic requirements of life. The substantial history of this approach shows that clear statements can be made about information processing bounds and how these influence the structure of sensory and behavioural systems and embodiment (Barlow, 1961; Brenner et al., 2000; Nehaniv et al., 2007; Pfeifer et al., 2007; Polani, 2009).

The information parsimony hypothesis has given rise to a body of research on the informational treatment of the perception-action loop of agents and the interactions with their environment. It has been shown that this can lead to global, fundamental insights in necessary bounds on behaviour (Polani et al., 2006), evolution of coordination (Sporns and Lungarella, 2006), intrinsic drives (Klyubin et al., 2008), successful search stramingoalinfotegies for tasks with sparse information (Vergassola et al., 2007), and behaviour structuring (van Dijk et al., 2009). These results are general in the sense that they do not require a specific model of brain mechanics. In this paper we will extend this previous work to the more specialised, though sufficiently general case of *goal-oriented* behaviour.

Goals

There are many cases, both in biological and in artificial settings, where the environment can be seen as offering rewards for certain types of behaviour. These rewards can range from as clear-cut as a treat given by a dog trainer to as diffuse as persistence. When such a reward measure is available to an agent, it can often be regarded as performing a certain task with an accompanying end-goal (Montague et al., 2004).

Although successful behaviour that appears goal-oriented is achieved, note that we do not want to imply that the organism or agent necessarily maintains an explicit representation of this goal. However, there is evidence for the case that human adults encode actions in terms of their outcomes (Hommel et al., 2001). Furthermore, brain structures have been located where activity is highly correlated to the goal of observed behaviour (Hamilton and Grafton, 2006), indicating an evolutionary drive towards goal-centred thought.

Moreover, recent research is beginning to show evidence for neural correlates of an individual's own goals, not limited to human brains, e.g. Saito et al. (2005); Spiers and Maguire (2006). Therefore we will adopt the viewpoint that certain behaviour, or in any case episodes of behaviour, can be seen as being driven by a concrete, identifiable goal.

Goal Information

We extend methods for informational treatment of the perception-action loop to explicitly include goal-directed behaviour. Here an agent needs to actively maintain information about its current goal. In the case of human beings it has been consistently argued that this is performed by the pre-frontal cortex (Montague et al., 2004). As any information processing this takes effort and consumes energy, thus, following the information parsimony hypothesis, it is expected that organisms attempt to optimise this process. Here therefore we study the necessary bounds of goal-information that has to be maintained at a given time. We show how these bounds can guide behaviour and that they can give rise to the emergence of certain behaviour properties, such as least-commitment planning, which traditionally is explicitly designed into computational approaches (Weld, 1994), and goal-concealing.

In the following two sections we will give a short introduction to concepts and notation used in this paper and an overview of the informational methods used to study the perception-action loop. Next, we introduce the main concept of the research presented here: *relevant goal information*. The effects of this quantity on behaviour and interpretations of these effects are then presented using a navigation-task example. Subsequently, we show how relevant goal information gives rise to a natural notion of *transition points*. Finally, we will relate our results to previous work and give a general discussion in the last section.

Concepts and Notation

When we talk about information, we refer to information in the information-theoretical formalism introduced by Shannon (1948). Here, the main elements are random variables, which we denote with capital letters, e.g. X. Such a variable can assume a specific value (small letter, x) from a given alphabet (curved capital, $\mathcal{X}$), subject to a probability distribution over the possible values: $\sum_{x\in\mathcal{X}} Pr(X = x) = 1$. To improve legibility we will, by abuse of notation, write $p(x)$ for both the entire distribution and for the probability that variable X assumes the value x, determined by the context. We use $p(x, y)$ and $p(y|x)$ for joint and conditional probabilities, respectively.

A probability distribution implies an 'uncertainty' about the value of a random variable. This uncertainty is quantified as the *entropy* $H(X) = -\sum_x p(x) \log p(x)$. We take 2 as the base of the logarithm, so that the unit of entropy is bits. Alternatively, the entropy can be seen as how much information on average is gained when learning the value of a random variable. The conditional entropy $H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x)$ determines the amount of uncertainty left about Y when the value of X is know.

The amount of information that on average is available both in X and Y can be calculated with the *mutual information* $I(X; Y)$. The mutual information can be defined as $I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$, which leads to the interpretation that it is the decrease in uncertainty about one variable when the value of the other one is known.

Finally. the expected value of a random variable is written as $\mathbb{E}[X]$, or $\mathbb{E}[X|\theta]$ when the value is conditioned on some parameters θ. The expected value is equal to the sum of the possible values, weighed by their probability: $\mathbb{E}[X] = \sum_x p(x)x$. Similarly, we can for instance write the conditional expected value of a function as $\mathbb{E}[f^\theta(X, y)|\theta] = \sum_x p(x|y, \theta) f^\theta(x, y)$.

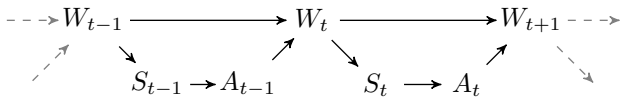
For a more elaborate background on the information-theoretical concepts and notations used in the current paper see Cover and Thomas (1991).

The Perception-Action Loop

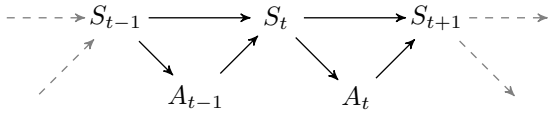
An agent is embodied and situated in an environment; it has direct contact to the environment through its sensors and actuators. Information abut the world is obtained through the sensors and influence the agent's actions, which in turn can affect the environment. This results in a *Perception-Action loop (PA-loop)* and, following Klyubin et al. (2004), we model this loop as a *causal Bayesian network* (CBN), as shown in Fig. 1(a). Such a network represents the relationship between the agent and the environment. At each time step t the agent perceives part of the state of the world w_t, resulting in a sensor state $s_t \in \mathcal{S}$. A fully reactive agent chooses its action $a_t \in \mathcal{A}$ based solely on this state. Its *policy* π defines the probability of performing these actions: $\pi(a_t|s_t) = p(a_t|s_t)$. When the agent performs an action, the world state is changed according to the *state transition probability distribution* $\mathcal{P}^{a_t}_{w_t,w_{t+1}} = p(w_{t+1}|w_t, a_t)$.

Without loss of generality, in the rest of this paper a simplified version of this model is used. It is assumed that the world is fully accessible to the agent, i.e. the sensor state reflects the full state of the world. For the CBN, this means that the world and sensor nodes can be collapsed, resulting in the network shown in Fig. 1(b). Consequently, we will use the term 'state' interchangeably for both world and sensor state.

As outlined in the introduction, we consider agents that operate in an environment that rewards certain behaviour. We are interested in how in this case the combined structure of the world and rewards can influence the structuring of behaviour. We assume that the reward that the agent receives is quantifiable. For instance, in a food-searching task the



(a) Complete Model



(b) Fully Accessible World

Figure 1: Causal Bayesian network of the perception-action loop, unrolled in time, showing (a) the complete model and (b) the case when the world is fully accessible.

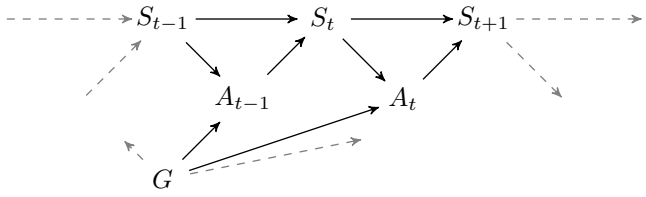


Figure 2: Causal Bayesian network of the perception-action loop, extended with the goal node.

agent can be presented a reward related to the nutritional value of the food when it is found. Another commonly used scheme is to represent the energy spent to perform a task as a penalty or negative reward for each time step that the goal is not reached. We will use the first model, as detailed further on.

These rewards are modelled by an *immediate-reward function* (Sutton et al., 1999) which gives the immediate reward that an agent will receive for performing action a_t when in state s_t and consequently finding itself in state s_{t+1}: $\mathcal{R}^{a_t}_{s_t,s_{t+1}} \in \mathbb{R}$. Given this function we can define the *state-action value function* (or *utility function*) $U^\pi(s_t, a_t)$ which gives the expected future reward of taking action a_t when in state s_t and subsequently following policy π (Sutton et al., 1999):

$$U^\pi(s_t, a_t) = \sum_{s_{t+1}} \mathcal{P}^{a_t}_{s_t,s_{t+1}} \left[\mathcal{R}^{a_t}_{s_t,s_{t+1}} + \gamma \mathbb{E}[U^\pi(s_{t+1}, A_{t+1})|\pi]\right], \quad (1)$$

where $\gamma \in [0, 1]$ is a discount factor to model preference for short term (low γ) or long term reward (high γ).

In this setting, a rational agent that performs goal-directed behaviour will try to gather as much reward as it can as fast as possible, effectively attempting to find an optimal policy π^* maximising the expected value of Eq. (1):

$$\pi^* = \arg\max_\pi \mathbb{E}\left[U^\pi(S_t, A_t)|\,\pi\right] \quad (2)$$

$$= \arg\max_\pi \sum_{s_t,a_t} p(s_t, a_t) U^\pi(s_t, a_t) \quad (3)$$

$$= \arg\max_\pi \sum_{s_t,a_t} \pi(a_t|s_t) p(s_t) U^\pi(s_t, a_t). \quad (4)$$

Information in the PA-Loop

With the formalisms outlined in the previous sections in place, we can look at the informational properties of the PA-loop. The arrows in the CBNs of Fig. 1 can be regarded as channels; the world 'transmits' information which the agent receives through its sensors and in turn the agent 'injects' information into the world through its actuators. The well established field of information theory then provides us with the tools to answer questions about the PA-loop in a concrete way in the terms of Shannon information (Shannon, 1948).

For instance, we can determine the amount of information that an agent on average takes in through its sensors to determine its actions using the mutual information between sensor states and actions $I(S_t; A_t)$. Not all information that is available in S_t is relevant to its current task and, following the hypothesis of information parsimony as discussed in the introduction, we assume that the agent will aim to minimise this quantity. The lower bound of the necessary amount of information intake to be able to achieve a certain level of utility can be quantified using the paradigm of *relevant information* (Polani et al., 2006).

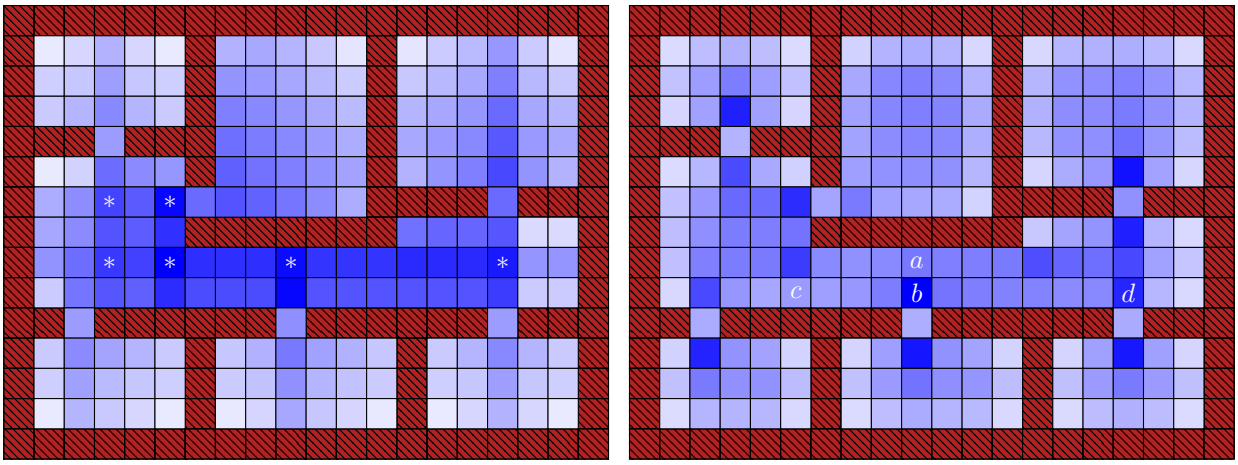
This is done by solving the following problem:

$$\min_{\pi(a_t|s_t)} \left[I(S_t; A_t) - \beta \mathbb{E}\left[U^\pi(S_t, A_t)|\pi\right]\right]. \quad (5)$$

The solution is a policy which minimises the state-information used to select actions while maximising the expected utility achieved by this policy. The parameter β can be varied to trade-off utility and information requirement; low β promotes information parsimony, high β puts more weight on utility. When β goes to infinity, the policy found will become optimal and the minimum amount of state information needed to act optimally is given by $I(S_t; A_t)$. As shown by Polani et al. (2006), the problem of (5) can be solved with an iterative algorithm that interleaves traditional algorithms of information theory (rate-distortion (Blahut, 1972)) and reinforcement learning (value iteration (Sutton and Barto, 1998)). This algorithm has the important property that the solution of (5) simultaneously fulfils Eq. (1).

Relevant Goal Information

The methods for relevant information are generally applicable to any case where a reward function can be defined. However, it is restricted to the analysis of a single task. Here we will extend the model of the PA-loop to enable us to handle an agent that could perform different tasks. To do so, we



(a) Relevant Goal Information (b) Goal Information Transitions

Figure 3: Grid world example for relevant goal information. Walls are denoted with a brown, hashed background. The remaining free cells comprise the set of states $\mathcal{S}$. The goal G is uniformly distributed and its alphabet $\mathcal{G}$ consists of the empty cells within the six rooms. The agent can perform four actions: move north, east, south or west. When such an action would move the agent to an occupied cell the action has no effect. The shading of the background of the free cells indicates (a) the total amount of relevant goal information for each cell and (b) the amount of new relevant goal information when arriving in a cell. Dark blue shading for high amount, light blue or white for low amounts The meaning of the asterisk and letter marks is explained in the text.

focus on the common case where this task can be determined by reaching a distinct goal. Here we do not discern how the current goal of an agent is selected; it can be imposed externally, such as a command given to a dog by its master, or it may be an intrinsically determined goal, as in the case of a hungry predator that decides to catch a certain prey. Instead, we only are concerned about the decision making process once a goal is given.

We introduce the new random variable G. The value of this variable, g, represents the current goal of an agent. Figure 2 shows how the CBN of the PA-loop is extended with this new variable. Note that we do not aim to study the case of an agent having several simultaneous goals. Rather, we concentrate on agents that select a specific goal from a discrete set of possible goals $\mathcal{G}$. After this selection the goal is fixed, until the goal is achieved or abandoned.

The new CBN shows that the policy now also depends on the current goal: $\pi(a_t|s_t, g) = p(a_t|s_t, g)$. Also, each separate goal gives rise to a distinct immediate reward function and thus to a separate goal-dependent utility function $U^\pi(s, g, a)$.

This extension of the model introduces an additional information source; apart from *sensory* information the agent now also needs to maintain and process *goal* information to guide its actions. Per the information parsimony hypothesis this is assumed to be costly and therefore we are interested in determining lower bounds on this amount of information needed to achieve a given performance. Analogous to the sensory case we term this the relevant *goal* information. In contrast, we will denote the traditional relevant information with relevant *sensory* information.

Whereas the relevant sensory information determines the minimum amount of sensory information necessary for a certain goal, we can also determine the minimum goal information necessary on average to achieve a certain utility, given the current state. By analogy to Eq. (5), this is done by solving the following minimisation problem:

$$\min_{\pi(a_t|s_t,g)} [I(G; A_t|S_t) - \beta\mathbb{E}[U(S_t, G, A_t)|\pi]] \qquad (6)$$

The solution to this problem, which is a policy trading off goal information parsimony with utility, controlled by the trade-off parameter β, can be found using the same iterative procedure used for relevant sensory information as described in (Polani et al., 2006).

As an example we use a navigation task in the grid world shown in Fig. 3(a). The set of states $\mathcal{S}$ consists of all unoccupied cells, the set of goals $\mathcal{G}$ contains the cells within the six rooms and the goal variable G is assumed to be uniformly distributed; any of the goals is as likely as another. The agent is rewarded when it achieves the current goal ($\mathcal{R}^{a_t}_{s_t,s_{t+1}} = 1$ if $s_{t+1} = g$, 0 otherwise) and a discount factor of $\gamma = 0.9$ is used.

As with relevant sensory information, we can study the trade-off between utility and relevant goal information by varying the value of β in Eq. (6). Figure 4 shows that the results of this trade-off are similar to that found for rele-

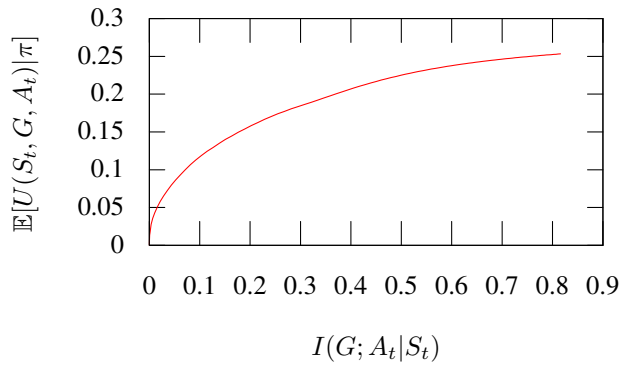


Figure 4: Trade-off between goal information (horizontal axis, bits) and expected utility (vertical axis).

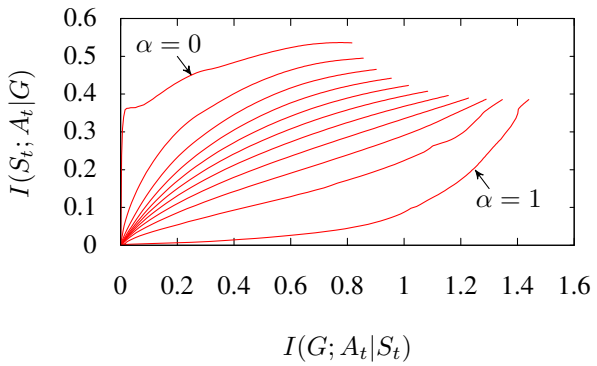


Figure 5: Trade-off between goal information (horizontal axis, bits) and sensory information (vertical axis, bits) for different values of $\alpha \in [0, 1]$, which controls preference for goal (low α) or sensory (high α) information parsimony.

vant sensory information; expected utility rises monotonically with higher goal information bandwidth, but the agent can still achieve a performance close to 90% of the maximum with as little as half of the optimal amount of information.

Besides utility, goal information may also have to be traded off against sensory information; a policy that minimises relevant goal information could require a higher average bandwidth for the sensors. We can combine equations (5) and (6) to take into account both costs:

$$\min_{\pi(a_t|s_t,g)} [(1-\alpha)I(G; A_t|S_t) + \alpha I(S_t; A_t|G) - \beta\mathbb{E}[U(S_t, G, A_t)|\pi]], \quad (7)$$

where α can be varied from 0 to 1 to reflect the relative cost of each process. Figure 5 shows that generally more relevant goal information is linked to an increase in sensory information, but that different weights result in different trade-offs.

We can extract the relevant goal information for each state separately, $I(G; A_t|s_t)$, as is shown in Fig. 3(a) for the policy achieving maximum expected utility. This example shows some interesting properties of relevant goal information. Firstly, in central states the agent tends to require more goal information than in more remote states or states close to walls. This is easily explained by the fact that in the central states the a priory probability of the direction the goal is in is roughly uniformly distributed; the goal can be on any side. When in the more distant states, however, the goal tends to be in a single direction. Only in exceptional cases does the agent need to deviate from going in this default direction and thus use extra goal information. Directly next to the walls the agent even only has to choose from the limited set of actions that do not make it run into a wall. Here the relevant goal information is bounded from above by the cardinality of this limited set. This also explains why the amount of relevant information in doorways is found to be often lower than in neighbouring states; here only two actions are useful.

Another observation is that local peaks in relevant goal information, marked with an asterisk in Fig. 3(a), can be found in front of doorways, even several cells away, most notably at 'crossing points' between different doorways. Trajectories of the agent tend to go from one of these peak cells to another. We will give an interpretation and explanation for this effect in the global discussion at the end of this paper.

Goal Information Transitions

In the example of the previous section we have only looked at single step scenarios. It shows that in different states the amount of goal information needed can vary. An interesting question is whether there is also a qualitative difference between the relevant goal information in different states. For instance, a bee flying out to search for food at first only has to consider which patch in its habitat is its target. Only when arrived at this patch it has to take into account the several individual resources (Bell, 1990). As another example, in our grid world, when the agent is in front of a doorway, it has to take into account whether the goal is in the neighbouring room or not. However, when it has just entered the room, this information is no longer relevant and it now has to focus on where exactly in the room the goal is. The model of relevant goal information given here can be used to analyse this development of goal information through time.

Given the single-step goal-information parsimonious policy as found in the previous section, we can determine how much of the relevant goal information in a certain state was not needed during the sequence leading to that state:

$$I(G; A_t|\mathbf{A}_0^{t-1}, s_t) = H(G|\mathbf{A}_0^{t-1}, s_t) - H(G|\mathbf{A}_0^{t}, s_t), \quad (8)$$

where $\mathbf{A}_0^t = (A_0, \ldots, A_t)$ denotes the sequence of actions from the start of the task to time step t. This amount of new relevant goal information is shown for our grid world case in Fig. 3(b), averaged over sequences of up to 5 time steps.

As one would expect, some of the cells where the total amount of relevant goal information is high (those marked in Fig. 3(a)) also stand out here; if in a cell more goal information is required than in the neighbouring cells, naturally a relatively high amount of this information is new. However, there are some notable differences: although the states where much new goal information is needed also require much total goal information, the opposite argument does not hold.

For instance, the cells marked a and b in Fig. 3(b) are shaded darkest in Fig. 3(a) and so require the most amount of information, with only a small difference between them. But there is a clear difference in how much of this information is new and different from the goal information that on average is required in the past before arriving in these cells. At cell b, in front of the doorway, the qualitative transition in goal information is much more pronounced. This same difference can be seen in the cells marked c and d; again, the total amount of relevant information for these cells is approximately the same, but for cell c more of this information is the same as already maintained by the agent in previous steps, showing a much less defined transition. All in all, we can note that the largest transitions are at doorways and at corners.

Discussion

Two Viewpoints

The result of minimisation of goal information is a policy where the agent often takes the same action, regardless of the goal; e.g. if going north works for all goals and going east only for a part of them the agent can always select going north and it can disregard all goal information. This leads to two complementary viewpoints for relevant goal information.

One is what we call the *least-commitment* (in the sense of least-commitment planning (Weld, 1994)) viewpoint. Because the actions taken by the agent are optimal for as many goals as possible, the amount of goals excluded by the actions are minimal. Although, in the methods described here, the goal does not change during a single run, because of the least-commitment property of the agent's policy, it will have a higher probability of still having behaved optimal if such a change does happen. The policy of the agent can be seen as keeping as many options open as possible. Thus, minimisation of relevant goal information causes the emergence of a least-commitment strategy.

This shows the relatedness of relevant goal information to *empowerment* (Klyubin et al., 2008). This quantity defines the maximum amount of possible observable control an agent has on its environment and is based on the same kind of informational treatment of the PA-loop as put forward in this paper. In a task-less setting empowerment leads to an intrinsic drive to least-commitment behaviour, whereas relevant goal information gives rise to such a drive in a goal-oriented agent.

The least-commitment viewpoint leads to the interpretation of states where relevant goal information is high as necessary decision points. If the goal can be in either of two rooms, the agent will not move towards one or the other until it has no other option. This occurs at the crossing points between doorways, where the agent has to make a decision and commit to one of the rooms.

Such an approach to delay decision making may not always be optimal, such as a driver who risks an accident by steering for a corner at the last moment at high speed. However, here these risks are assumed to be contained in the reward function, rendering such policies suboptimal and thus no longer considered by the agent.

Another interpretation arises from the *goal-concealing* viewpoint. This viewpoint is obtained by noting that the mutual information between goal and action can not only be seen as how much goal information is needed to decide on an action, or how much information the goal gives about the action, but also how much information the actions give about the goal (a similar viewpoint for sensory relevant information is taken by Salge and Polani (2009)). This means that by minimising relevant goal information the agent gives away as little information as possible about its goal to an external observer. This observer could see this as the emergence of a goal-hiding strategy.

From this viewpoint the peaks in relevant goal information at crossing points can be explained by noting that the actions taken here give away a lot of information about the goal of the agent. When at a crossing point between two rooms, the observer does not know in which room the goal is, but after seeing the action he can exclude all the cells in the room the agent moved away from.

Sub-Goals

In the field of Reinforcement Learning (RL) there has been a lot of recent activity on the subject of higher level behaviour structuring, task decomposition and automatic sub-goal discovery (Barto and Mahadevan, 2003). A large amount of algorithms for automatic behaviour structuring have resulted from this. For instance, the intuition that so called 'bottleneck' or 'funnel' states in an environment, such as doorways, are salient sub-goals has led to methods being developed based on visitation count (McGovern and Barto, 2001; Kretchmar et al., 2003; Asadi and Huber, 2005) and graph-theoretical techniques (Şimşek et al., 2005; Kazemitabar and Beigy, 2009; Şimşek and Barto, 2009). Other approaches that are also based on assumptions about the structure of the world, but using less strict definitions of what may constitute a 'good' sub-goal, include state space segmentation/clustering (Bakker and Schmidhuber, 2004; Mannor et al., 2004), relative novelty (Şimşek and Barto, 2004), sensation/action co-occurrence (Digney, 1996) or transitions

(Hengst, 2002; Kozlova et al., 2009), causal-graph decomposition (Jonsson and Barto, 2006) and the use of data-mining techniques (Kheradmandian and Rahmati, 2009). Finally, a separate class of algorithms does not focus on structure of goals, but on segmentation, clustering and abstracting common state-action sequences (Sun and Sessions, 2000; Pickett and Barto, 2002; Girgin et al., 2006).

All these methods indicate their usefulness by showing increased learning performance in certain RL tasks. Also, they show that skill transfer, made possible by task segmentation, can be highly beneficial (Perkins and Precup, 1999; Konidaris and Barto, 2007). However, hardly any comparison of the performance of different approaches has yet been done. This is not surprising, since the methods can differ greatly and, more importantly, they are based on different, designer imposed, assumptions about what is a good way to structure a task. In these papers the structural properties of a sub-goal or sub-task are defined for a particular domain of interest, after which a solution is engineered for these specific properties.

The results of the current paper, however, suggest a more fundamental, biologically/Alife motivated definition of sub-goals: a sub-goal is achieved when a significant qualitative change of the task at hand occurs, which occurs when the actions of an agent are guided by a new component of, or new information about, the goal not taken into account earlier. As shown earlier, the notion of relevant goal information can be used to identify such transitions. Note that the informational treatment of the PA-loop is independent of domain, architecture and particular implementations and therefore we do not need any of the assumptions made in the engineering solutions. The biologically plausible hypothesis of information parsimony is sufficient for the treatment of emergence of sub-goals.

References

Asadi, M. and Huber, M. (2005). Accelerating Action Dependent Hierarchical Reinforcement Learning Through Autonomous Subgoal Discovery. In *Proceedings of the ICML 2005 Workshop on Rich Representations for Reinforcement Learning*.

Bakker, B. and Schmidhuber, J. (2004). Hierarchical Reinforcement Learning Based on Subgoal Discovery and Subpolicy Specialization. In *Proceedings of the 8-th Conference on Intelligent Autonomous Systems, IAS-8*, pages 438–445.

Barlow, H. B. (1961). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W., editor, *Sensory Communication*, chapter 13, pages 217–234. MIT Press, Cambridge, MA.

Barto, A. G. and Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, 13(1-2):41–77.

Bell, W. J. (1990). Searching Behavior Patterns in Insects. *Annual Review of Entomology*, 35:447–467.

Blahut, R. E. (1972). Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Transactions on Information Theory*, 18:460–473.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R. (2000). Adaptive Rescaling Maximizes Information Transmission. *Neuron*, 26(3):695–702.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory*. Wiley-Interscience, New York, NY, USA.

Digney, B. (1996). Emergent Hierarchical Control Structures: Learning Reactive / Hierarchical Relationships in Reinforcement Environments. In *Proceedings of the Fourth Conference on the Simulation of Adaptive Behavior: SAB 98*, pages 363–372. MIT Press.

Girgin, S., Polat, F., and Alhajj, R. (2006). Learning by Automatic Option Discovery from Conditionally Terminating Sequences. In *ECAI*, pages 494–498.

Hamilton, A. F. d. C. and Grafton, S. T. (2006). Goal Representation in Human Anterior Intraparietal Sulcus. *Journal of Neuroscience*, 26(4):1133–1137.

Hengst, B. (2002). Discovering Hierarchy in Reinforcement Learning with HEXQ. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 243–250, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, 24(05):849–878.

Jeffery, W. (2001). Cavefish as a Model System in Evolutionary Developmental Biology. *Developmental Biology*, 231(1):1–12.

Jonsson, A. and Barto, A. (2006). Causal Graph Based Decomposition of Factored MDPs. *Journal Of Machine Learning Research*, 7:2259–2301.

Kazemitabar, S. J. and Beigy, H. (2009). Automatic Discovery of Subgoals in Reinforcement Learning Using Strongly Connected Components. *Advances in Neuro-Information Processing*, 5506:829–834.

Kheradmandian, G. and Rahmati, M. (2009). Automatic Abstraction in Reinforcement Learning Using Data Mining Techniques. *Robotics and Autonomous Systems*, 57(11):1119–1128.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004). Tracking Information Flow through the Environment: Simple Cases of Stigmergy. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R. A., editors, *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 563–568. The MIT Press.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2008). Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLoS ONE*, 3(12):e4018.

Konidaris, G. and Barto, A. (2007). Building Portable Options: Skill Transfer in Reinforcement Learning. In *IJCAI'07: Proceedings of the 20th International Joint Conference on Artifical intelligence*, pages 895–900, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kozlova, O., Sigaud, O., and Meyer, C. (2009). Automated Discovery of Options in Factored Reinforcement Learning. In *Proceedings of the ICML/UAI/COLT Workshop on Abstraction in Reinforcement Learning*, pages 24–29, Montreal, Canada.

Kretchmar, R., Feil, T., and Bansal, R. (2003). Improved Automatic Discovery of Subgoals for Options in Hierarchical Reinforcement Learning. *Journal of Computer Science & Technology*, 3:9–14.

Laughlin, S. B., de Ruyter van Steveninck, R. R., and Anderson, J. C. (1998). The Metabolic Cost of Neural Information. *Nature Neuroscience*, 1:36–41.

Mannor, S., Menache, I., Hoze, A., and Klein, U. (2004). Dynamic Abstraction in Reinforcement Learning via Clustering. In *ICML '04: Proceedings of the twenty-first International Conference on Machine Learning*, page 71, New York, NY, USA. ACM.

McGovern, A. and Barto, A. G. (2001). Automatic Discovery of Subgoals in Reinforcement Learning using Diverse Density. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 361–368, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Montague, P. R., Hyman, S. E., and Cohen, J. D. (2004). Computational Roles for Dopamine in Behavioural Control. *Nature*, 431(7010):760–7.

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

Perkins, T. and Precup, D. (1999). Using Options for Knowledge Transfer in Reinforcement Learning. Technical report, University of Massachusetts, Amherst, MA, USA. UM-CS-1999-034.

Pfeifer, R., Lungarella, M., Sporns, O., and Kuniyoshi, Y. (2007). On the Information-Theoretic Implications of Embodiment – Principles and Methods. In *Proc. of the 50th Anniversary Summit of Artificial Intelligence*, volume 4850, pages 76–86. Springer-Verlag.

Pickett, M. and Barto, A. G. (2002). PolicyBlocks: An Algorithm for Creating Useful Macro-Actions in Reinforcement Learning. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 506–513, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Polani, D. (2009). Information: Currency of Life? *HFSP Journal*, 3:307–316.

Polani, D., Nehaniv, C. L., Martinetz, T., and Kim, J. T. (2006). 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*.

Saito, N., Mushiake, H., Sakamoto, K., Itoyama, Y., and Tanji, J. (2005). Representation of Immediate and Final Behavioral Goals in the Monkey Prefrontal Cortex During an Instructed Delay Period. *Cerebral Cortex*, 15(10):1535–46.

Salge, C. and Polani, D. (2009). Information Theoretic Incentives for Social Interaction . Technical Report 495. presented at ECAL Workshop on Organisation, Cooperation and Emergence in Social Learning Agents.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656.

Şimşek, Ö. and Barto, A. G. (2004). Using Relative Novelty to Identify Useful Temporal Abstractions in Reinforcement Learning. In *ICML '04: Proceedings of the Twenty-First International Conference on Machine Learning*, page 95, New York, NY, USA. ACM.

Şimşek, Ö. and Barto, A. G. (2009). Skill Characterization Based on Betweenness. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 1497–1504.

Şimşek, Ö., Wolfe, A. P., and Barto, A. G. (2005). Identifying Useful Subgoals in Reinforcement Learning by Local Graph Partitioning. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, pages 816–823.

Spiers, H. J. and Maguire, E. A. (2006). Thoughts, Behaviour, and Brain Dynamics During Navigation in the Real World. *Neuroimage*, 31(4):1826–40.

Sporns, O. and Lungarella, M. (2006). Evolving Coordinated Behavior by Maximizing Information Structure. In *Artificial Life X*, pages 323–329. MIT Press.

Sun, R. and Sessions, C. (2000). Self-Segmentation of Sequences: Automatic Formation of Hierarchies of Sequential Behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 30(3):403–418.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between MDPs and Semi-MDPs: a Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112(1-2):181–211.

van Dijk, S. G., Polani, D., and Nehaniv, C. L. (2009). Hierarchical Behaviours: Getting the Most Bang for your Bit. In *Proceedings of the 10th European Conference on Artificial Life*. Springer (In Press).

Vergassola, M., Villermaux, E., and Shraiman, B. I. (2007). 'Infotaxis' as a Strategy for Searching Without Gradients. *Nature*, 445(7126):406–9.

Weld, D. S. (1994). An Introduction to Least Commitment Planning. *AI Magazine*, 15(4):27–61.