Advances in Complex Systems

Informational Organization of Task Structure

SANDER G. VAN DIJK
Adaptive Systems Group, University of Hertfordshire, College Lane
Hatfield, Hertfordshire, United Kingdom
s.vandijk@herts.ac.uk

DANIEL POLANI
Adaptive Systems Group, University of Hertfordshire, College Lane
Hatfield, Hertfordshire, United Kingdom
d.polani@herts.ac.uk

Received (received date)
Revised (revised date)

We consider the effect of the structure of an agent's cognitive system (its 'embrainment') on the organization of its behavior, most notably from the viewpoint of constraints that this structure may impose on the agent's information processing capabilities. In doing so, we outline the basis of a unified information-theoretic framework to treat the internal organization of decision making for an agent guided by a set of tasks and by the relevance of available information. Using this framework, we show several examples of how starting from the intrinsic considerations of limiting embrainment uncovers a rich spectrum of structure in environment and tasks, which traditional approaches need to specify a-priori, such as salient transition points, task similarity, and local and global organization of the environment. We argue that this structure, which in our approach is inherently relevant to the behavior of an agent, could supply an important guide for the self-organization of an agent charged with a set of tasks.

Keywords: embrainment, self-organization, informational structure; relevant information; sub-goals; task clustering

1. Introduction

When considering an agent charged with a task, or a set of possible tasks, one of the most important aspects is clearly the degree in which the agent is able to perform its task(s). We marvel at how biological agents seem to do the things they do perfectly, especially since much of their behavior is self-organized. In the creation of artificial intelligence we also strive to get agents to do the things we want them to do as good as possible, but to accomplish this a designer often needs to supply a high level of pre-structuring to the agent. Here, we present a unified information-theoretic framework as a step towards self-organization guided by necessary structure of an agent's decision making facilities.

Over the last few decades it has become clear that an agent's *embodiment*, meaning its capabilities to interact with the environment, is crucial to its performance. Although in hindsight it may not seem surprising, a large paradigm shift was needed to see that if an agent's embodiment fits nicely with the structure of its environment and its goals, it will perform better. In this paper we will extend, and *invert*, this intuition, to the agent's *internal* decision making facilities, which one could call *embrainment*: the organization and structure of an agent's decision making apparatus ('brain') influences the agent's capabilities, and an agent will generally do well to organize its decision making process in such a way to utilize whatever structure is available in its environment and in its tasks. Examples of such structure can be the distinctions and commonalities of different areas of the environment or transition points between them, as well as similarities between tasks. The benefit of exploiting such structure is most obvious when an agent has to operate under restrictions of its embrainment. However, also in cases where cognitive constraints are not directly an issue, alleviating cognitive burden can still be a good thing to do in light of efficient use of resources, as has been hypothesized in the context of cognitive load optimization in organisms [11]. Also, as we will show, analyzing the minimal properties of different facets of an agent's decision making process can uncover structure that could be useful to an agent with unlimited embrainment, such as when learning new tasks.

In this paper we will discuss these aspects in the light of information-theoretic treatment of the internal organization of decision making for an agent guided by a set of tasks and by the relevance of available information. The use of information-theoretic methods to analyze and guide the behavior of agents is becoming increasingly popular. This paradigm is based on the view of an agent as an information processing system that is interacting with the environment through a sensory and an actuation channel [26]. As we will show below, concepts and tools from the field of *Information Theory (IT)* are well suited to model and analyze such systems. This kind of modeling can lead to fundamental insights, such as strict limits on control [27], the best performance that an agent can achieve under sensory limitations [16, 25], and how embodiment induces information structure in sensory inputs [14]. Considering these insights, it is hypothesized that optimizing a biological organism's information intake, processing, and output has an evolutionary advantage [15]. This hypothesis has led to quantitative models which allowed the development of a number of methods for self-organization of behavior, achieving coordinated behavior by maximizing informational quantities or properties, such as informational structure [21], information about the location of a target [30] (which in a multi-agent setting can be sufficient to induce flocking behavior [18]), observable control [10], or predictability of future information [3, 1].

In this paper, we will add to this type of purely intrinsic self-organization the requirement to solve a set of tasks. With this we mean that we are interested in agents equipped with such a set of tasks and their respective measures of performance (i.e. a *multi-task* scenario), and a drive to maximize their performance on these tasks. While many methods have been constructed to achieve optimal solutions for such scenarios, our approach is different by taking into account an agent's possibly limited embrainment. By doing so we aim to reconstruct *solely from intrinsic considerations* the concepts and structure that are commonly used as an external guide in other approaches. We will model the tasks in the well known framework of *Markov Decision Processes (MDPs)*, as we will show in Sec. 2. This framework however only deals with optimality of behavior; the traditional MDP framework ignores the trade-off between optimality and cognitive burden, bandwidth constraints in decision making facilities, and structure other than that in the performance measures. We will incorporate these aspects with a unified, information-theoretical extension to the classical MDP framework, as described in Sect. 3. In the remainder of the paper we will then use this framework to see how an agent can uncover structure in the environment and the task-set, and finally discuss how an agent could utilize these results to organize its decision process.

2. Multi-Task Scenarios as Families of MDPs

In this section we will discuss the concepts that formalize the main guiding drive of our agents: to perform optimally on a set of tasks. We formulate a multi-task scenario as a family of so called *Markov Decision Processes (MDPs)*, similar to how *Multi-Task Learning* problems are defined in [23]. Each task is a single MDP, in which the possible states of the system and the actions available to the agent are described by some sets $\mathcal{S}$ and $\mathcal{A}$. The effects of actions on the state of the world are modeled by a transition probability distribution function, $P^{s_{t+1}}_{s_t,a_t} = p(s_{t+1}|s_t, a_t)$. A reward function $R^{s_{t+1}}_{s_t,a_t}$ gives the immediate reward r_t that an agent will receive at each transition. Commonly, this reward function defines the goal of a task as a certain world state that should be reached, for instance by giving a high reward when a goal state is reached, or, as we will use in this paper, by giving a negative reward/cost of -1 for every step in which the goal state is not reached. In the family of MDPs that we will consider, the current goal is selected from some set, $\mathcal{G}$, according to the task probability distribution $p(g)$. For each task, the state and action sets $\mathcal{S}$ and $\mathcal{A}$ and the transition distribution $P^{s_{t+1}}_{s_t,a_t}$ are the same, however each goal g is associated with its own unique reward function $R^{s_{t+1}}_{g,s_t,a_t}$. Finally, at each time step, the agent selects its action based on the current state and task according to the distribution determined by its *policy* $\pi(a_t|s_t, g) = p(a_t|s_t, g)$.

Given this formulation, the performance of an agent can be measured by the total reward it gathers when performing the tasks. The expected value of this measurement for a single task for all start states and following a certain policy π is given by the *value function* $V^{\pi}_{g}(s_t)$, with which one can determine the *utility* $U^{\pi}_{g}(s_t, a_t)$

of taking an action in each state during each task:

$$\begin{aligned} V_g^\pi(s_t) &= E[r_t + r_{t+1} + r_{t+2} + \dots] \\ &= \sum_{a_t \in \mathcal{A}} \pi(a_t|s_t, g) \sum_{s_{t+1} \in \mathcal{S}} p(s_{t+1}|s_t, a_t)\Big(R^{s_{t+1}}_{g,s_t,a_t} + V_g^\pi(s_{t+1})\Big) \qquad (1) \\ U_g^\pi(s_t, a_t) &= \sum_{s_{t+1}} p(s_{t+1}|s_t, a_t)\Big(R^{s_{t+1}}_{g,s_t,a_t} + V_g^\pi(s_{t+1})\Big) \qquad (2) \end{aligned}$$

A policy is said to be optimal when it maximizes the expected utility over all tasks and states, $E[U_G^\pi[S_t, A_t]$. The field of *Reinforcement Learning (RL)* attempts to construct methods for finding such a policy. Great progress has been made in this area, however, there are several implicit assumptions underlying the commonly used traditional formulation described above. For instance, it is assumed that the agent's information intake and processing pathways have unlimited bandwidths that make all possible policies feasible. For a range of problems this may be the case, however, for many real world problems this may not be the so. A formal framework for treating possible informational constraints could give important insights into the properties and bounds of those constraints and their effect on performance. In the following sections we will outline the basis of such an informational framework that can treat this kind of restrictions in a unified way and without the need of further assumptions, and apply it to multi-task scenarios.

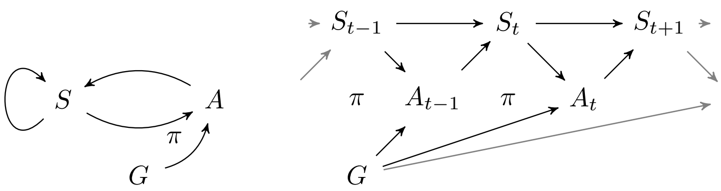
3. Informational Framework

Due to the probabilistic nature of state transitions and the agent's policy as given above, the state and selected action at a certain time and the selected task can be treated as random variables, denoted by S_t, A_t, and G (without a time index, as the goal persists over time). Their interactions can be visualized informally as the *Perception-Action loop (PA-loop)* of Fig. 1a, and unrolling this loop gives the formal description of the *Causal Baysian Network (CBN)* as shown in Fig. 1b [27, 9].

The edges of the network show the pathways of information in the system: information about both the goal and the current state is used to select an action, the execution of which transfers information into the world by influencing its future state. Using the CBN and the information-theoretic tools of the following sections we shall analyze these pathways in the remainder of this paper.

3.1. *Information*

Again, we are interested in the effect of cognitive burden and possible constraints on this burden on the behavior and performance of an agent. As mentioned in the introduction, an agent can be seen as an information processing system, and in this view cognitive burden can be correlated with the amount of information that the system needs to process, and/or the bandwidth requirements that need to be



(a) (b)

Fig. 1: (a) Perception-action loop (Arrows in (a) are to be understood informally as scheme of influence between the different variables in the system, while the arrows in (b) are to be understood in the formal framework of Bayesian Networks.(?)) (b) Causal Bayesian network of the perception-action loop unrolled in time. The state of the world at time t is denoted by S_t, the action selected by the agent, according to its policy π, by A_t. The current task is determined by G, which persists throughout time.

fulfilled, to achieve an agent's behavior. Such notions can be formalized by applying concepts and methods from the field of Information Theory directly to the MDP and Bayesian models described above. In this section we will give a short introduction to this field and its concepts, to construct the basic understanding needed for our main work. For a more elaborate introduction to IT see for instance [7].

The amount of Shannon information one gains on average when one learns the value of a random variable X is expressed by its *entropy* $H(X) = -\sum_x p(x) \log p(x)$, and the *conditional entropy* $H(Y|X)$ gives the average amount of information left to be learned about Y when we already know the value of X. The average amount of information shared by two variables X and Y, or, equivalently, the amount of information that the value of one variable on average gives about the other, is measured by the *mutual information* $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(Y) - H(Y|X)$. The *conditional mutual information* $I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$ gives the amount of information shared by X and Y beyond that what is already given by Z. The unit of these quantities is determined by the base of the logarithm used; we will use base 2 and thus measure all quantities in *bits*. Information is symmetric, meaning $I(X;Y) = I(Y;X)$, note however that a CBN induces directionality, and thus specifies which variables receives information from which other variable(s). The use of information theory in conjunction with a CBN allows one to model how information is transferred between agent and environment (embodiment) and how its processing inside the agent is organized (embrainment).

The amount of information available separately in the world state and in the task description are given by the resolution of the entropies $H(S_t)$ and $H(G)$ respectively. Generally the two variables will be correlated, since a state distribution will be formed from navigating to a specific goal. This means that the total amount of information, which is equal to the joint entropy $H(S_t; G)$, will be less, i.e. $H(S_t; G) = H(G) + H(S_t|G) < H(G) + H(S_t)$. However, for simplicity we will assume in our examples that the world state is independent from the current goal, i.e. $H(S_t|G) = H(S_t)$, as is the case at the start of each navigation task.

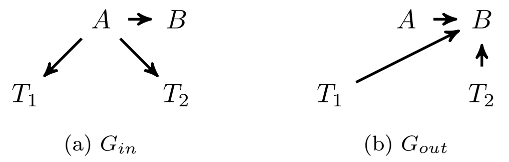
Typically, one can expect that an agent uses only a fraction of the available information at any time to select its action. Because of our assumption of independence between S_t and G, this fraction can be partitioned in $I(S_t; A_t|G)$, the information taken uniquely from the world state, and $I(G; A_t|S_t)$, the unique amount about the goal. Note that this independence does not imply that $I(S_t; A_t|G) = I(S_t; A_t)$ or $I(G; A_t|S_t) = I(G; A_t)$. As we will also discuss in a later section, there is a strong relationship between state and goal such that one of them is not very informative without knowledge of the other one.

The quantities listed here are fully determined by the world dynamics, as modeled by $P^{s_{t+1}}_{s_t,a_t}$, and the agent's dynamics given by π, and impose strict lower bounds on the agent's action selection facilities. For instance, given a world and a policy, the resulting value of $I(S_t; A_t|G)$ gives a lower bound for the sensory bandwidth of the agent: the agent *needs* to be able to take in this amount of information on average per step, or it *will not* be able to execute the respective policy correctly, *regardless of how this policy is implemented* [16]. The same holds for $I(G; A_t|S_t)$: if some constraint on the embrainment of the agent causes it to not have access to this amount of information about the goal, the policy is not feasible. We will elaborate on this in Sec. 3.3, as well as on the problem of selecting a policy that minimizes these informational requirements to alleviate these restrictions.

3.2. *Information Bottleneck*

As we express all constraints of the embrainment in information-theoretic terms, we will need to make use of an inherently information-theoretic concept to treat such constraints and study the structure of information passed through a limited channel. This concept is the *Information Bottleneck (IB)*, introduced in [24]. It can be used to extract the information in one signal that is relevant to another. For instance, imagine we have a random variable Y of which the value is determined by another random variable X according to a conditional probability distribution $p(y|x)$. If we are interested in what information in X is relevant to the value of Y, we can introduce another variable, X', and try to find a probabilistic mapping $p(x'|x)$, such that it minimizes the amount of information captured about X in the construction of X', while still retaining at least a certain amount of information that is relevant to Y. The IB method, named this way because X' can be seen as a *bottleneck variable* through which information from X about Y is squeezed, achieves exactly this.

In the rest of this paper we will use the more general *Multivariate Information Bottleneck (MIB)* methods for more complex Bayesian networks consisting of more



(a) G_{in} (b) G_{out}

Fig. 2: Example of a parallel information bottleneck, where two independent bottleneck variables T_1 and T_2 capture information from A that is relevant to B.

than just two variables [20]. In this method, one constructs two Bayesian networks: G_{in}, describing which variables should capture information from which other variables, and G_{out}, describing which variables determine the relevance of information. Figure 2 shows a so called *parallel* information bottleneck as an example of an MIB where two bottleneck variables are used in parallel to compress information about A that is relevant to B. One then tries to find an assignment of the bottleneck variables consistent with G_{in} that minimizes the information shared between all variables and their parents in G_{in}, under a constraint on the minimum amount of information shared between the variables and their parents in G_{out}. Using the method of Lagrange multipliers, this leads to the problem of minimizing the following Lagrange equation:

$$\Lambda\Big(p(\mathbf{T}|\mathbf{Pa}_{\mathbf{T}}^{G_{in}}),\beta\Big) = \sum_i I(X_i;\mathbf{Pa}_{X_i}^{G_{in}}) - \beta\sum_i I(X_i;\mathbf{Pa}_{X_i}^{G_{in}}), \tag{3}$$

where the minimization is over all possible conditional distributions of the list of bottleneck variables $\mathbf{T}$, the sums are over all variables in the networks, and $\mathbf{Pa}_X^G$ denotes the list of parents of variable X in graph G. The non-negative Lagrange multiplier β can be set to determine a trade-off between compression and predictive power: low β favors compression of the input variable(s) (A in the example of Fig. 2), whereas when β grows towards infinity the constraint of capturing all relevant information about the target variable (B in Fig. 2) is hardened. Several methods exist to solve this problem [20].

3.3. *Relevant Information*

Whereas the IB methods are used to extract and study the information relevant to a given random variable, for an agent the major indicator of relevance is its performance. It needs to extract and access that information which is critical to the utility of its actions; failure in doing so limits the set of feasible policies and could thus negatively affect its performance. Given a certain bandwidth, there will be an upper bound on the achievable performance, and, vice versa, given a desired level of performance there will be a lower bound on the bandwidth needed to achieve that level. As discussed in the introduction, there are several arguments for a drive in the agent to operate on the optimal trade-off between these factors, where it achieves the best performance for a given level of bandwidth, or, equivalently, where it achieves the smallest bandwidth requirements for a given level of performance. This idea is formulated by the notion of *Relevant Information (RI)*[16], which gives a way of finding this optimal trade-off and a policy that achieves it.

Each policy determines a certain value of $I(S_t; A_t|G)$, $I(G; A_t|S_t)$, and $E[U^\pi_G(S_t, A_t)]$, giving the minimum bandwidth of the agent's sensory and task channels required to carry out such a policy and the performance level of that policy. The optimal trade-off is found by maximizing the performance over all policies under fixed bandwidth, or, again equivalently, by minimizing the bandwidth under fixed performance. This is similar to the IB, with that difference that the agent's performance, i.e. the utility of actions, takes on the role as indicator of relevance, and using similar methods we can formulate a minimization problem to find a trade-off. For instance, if there is a drive to minimize the sensory bandwidth we arrive at the minimization of the Lagrange equation

$$\Lambda\Big(\pi(a_t|s_t, g), \beta\Big) = I(S_t; A_t|G) - \beta E[U^\pi_G(S_t, A_t)] \tag{4}$$

Note that this problem has a form similar to the classic Rate-Distortion problem, with the important difference that here the utility is not an a-priori given distortion measure, but depends itself on the policy. Therefore the method used to solve this problem extends the iterations of the classic Blahut-Arimoto algorithm for Rate-Distortion [4] to ensure that the utility is consistent with the policy at each step [16]. The self-consistent solutions of these problems that are used in this iterative methods, as well as the solutions for the other similar problems in this paper, can be found in Appendix B.

As with the IB, the Lagrange multiplier β gives us the possibility to determine the full range of possible trade-offs: when β goes to zero the information intake will go to zero, resulting in the best policy for a 'blind' agent, whereas increasing β puts more emphasis on optimality and less on informational parsimony. When β approaches infinity, performance gets close to optimal. The value of $I(S; A|G)$ acquired at the minimum of (4) is referred to as the *relevant sensory information*, because this is the amount of information that is actually relevant to achieving a certain performance. Any information beyond that can be discarded. It is important to note that this is an invariant property of the agent-environment system, in the same sense as for instance how in physics the minimum amount of energy needed to raise an object in a potential field is invariant: an agent that is not capable to take in and process at least the relevant information will *in no way* be able to achieve the required performance. An equivalent argumentation holds for the *relevant goal information*[16], which is the minimum required information about the current task, measured by $I(G; A_t|S_t)$, needed to achieve a certain level of performance (utility). Optimization of this latter quantity has as a result that the agent tries not to have to make a distinction between different goals as much as possible, for which goal information is needed. This is achieved by a policy that prefers to take actions that are good for as many goals as possible, and thereby performing a policy that seems

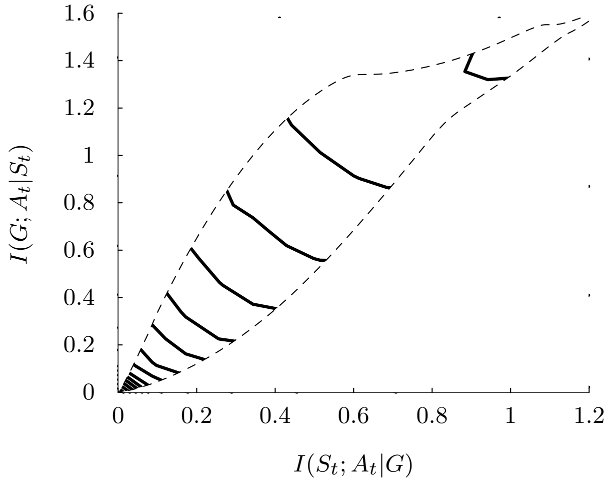


Fig. 3: Trade-offs between sensory information, goal information, and utility, obtained for a 6-room navigation scenario. The solid lines show trade-offs between sensory and goal informtion that achieve the same level of performance. The dashed lines mark the boundaries of the sensori-motor optimized area, in which all optimal state-/goal-information/performance trade-offs lie.

to keep its options open or hide its intentions [29].

State information and goal information interact with each other, even under our assumption that the current goal and state are independent. To see this, imagine that you are in a strange city and want to meet a friend for dinner. Knowing the location of the agreed upon restaurant does not help you decide which way to go if you are lost and do not know where you are. Likewise, information about where you are is not very usefull when you do not know which restaurant to go to. This property where two or more sources together give more information about other variables beyond the sum of the information they give seperately is called *interaction information* [12], or *synergy* [8], and can in our framework be seen by the fact that in non-trivial cases $I(S_t;A_t) \ll I(S_t;A_t|G)$ and $I(G;A_t) \ll I(G;A_t|S_t)$.

Another interesting facet of the interplay between state and task information is that one can be traded off for the other, without affecting an agent's performance [29]. This means that if there is a restricted bandwidth on one, or if processing one is relatively more costly, an agent could overcome this by employing more of the other type of information. To show this, we can apply the RI methods to a total informational cost, consisting of the weighted sum of sensory and goal information, $\alpha I(S_t;A_t|G) + (1-\alpha)I(G;A_t|S_t)$, where we can choose $\alpha \in (0-1)$ to trace out the optimal trade-off for all possible relative costs of the two types of information.

If we solve this problem over the possible ranges for both α and β, we can visualize the possible trade-offs for different levels of performance. Figure 3 shows for instance the results for the 6-room navigation scenario of Fig. 6. Moreover, we obtain the full range of *sensori-motor optimized behavior*: any policy outside of this set, marked by the area within the dashed lines in Fig. 3, is suboptimal for *any* given bandwidths-performance combination, since there is always another policy that achieves better performance with the same amount of information, and/or decreases the information bandwidth requirements on one or both of the sources of information without deteriorating performance.

In the remainder of this paper we will demonstrate how the framework as set out in this section can be extended to uncover a rich organization of environment and task structure.

4. From Quantification to Qualification

The methods of the previous section give a quantification of informational trade-offs. But we can go beyond that: we can also analyze the qualitative properties of these trade-offs and the policies that achieve these trade-offs. We have already briefly mentioned the example of identifying salient transitions points, on which we will elaborate here.

To do so, let us reconsider the necessary structure of the information processing facilities ('embrainment') of an agent. The agent needs to have a long-term memory that stores the full amount of goal information, which it needs to access at each decision step to select the proper action. One can imagine that there is some overlap in the type of information that is needed at consecutive time-steps. For example, when an agent is navigating through a room towards the next room, it reuses information about its relative position to the doorway leading to the target room at each step, until it enters the next room. This recollection of the same information over and over can be seen as unnecessary cognitive burden, which could be overcome by extending the agent's cognitive structure with a short-term working memory. In this working memory the agent could maintain the immediate goal information throughout the steps for which it is relevant, and alleviate the bandwidth requirements on the long-term memory by only 'downloading' the bare minimum of additional information from this memory when necessary. Naturally, this assumes that access to such a working memory is cheaper for the decision making center than access to the long-term memory. One can draw a parallel between this kind of working memory and the fast caching memory of a modern CPU that is used to alleviate bandwidth constraints on a computer's RAM.

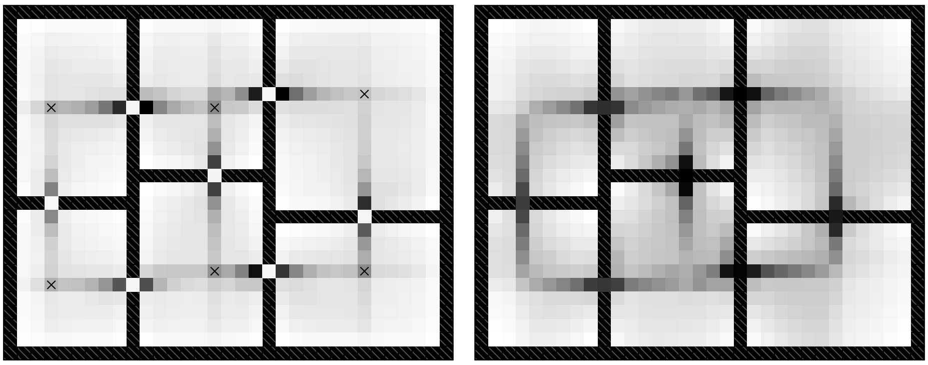
These considerations raise the question of what the minimum is of the amount of information that needs to be transferred into such a working memory from the long-term goal memory. This requirement is determined by the amount of goal information needed at this time that is not available in the information that is possibly already in the working memory, which is all the information downloaded and used to select actions up until the previous time step. To clarify, imagine for example that the agent wants to travel from one room to another, connected by a hallway. When entering the hallway, the relevant information about the final goal is only to make the distinction of on which end of the hallway the second room is. The agent ‘downloads’ this information into its working memory once, after which it is maintained and used for deciding where to walk to while traversing the hallway. Finally, when the agent enters the second room it needs new information that was not used earlier: “Where in the room did I want to go to?”

So, we want to quantify the amount of goal information that is relevant to the current action selection, beyond that which was relevant to past decisions and therefore possibly already available in the working memory. These past decisions are reflected by the state-action pairs encountered thus far, which we will denote as the random variable $\mathbf{E}_{t-1} = (S_0, A_0, S_1, A_1, \dots, S_{t-1}, A_{t-1})$. The *new* task information that is needed on average when an agent encounters a specific state s_t is then equal to $I(G; A_t|\mathbf{E}_{t-1}, s_t)$. In previous work we presented estimations of this quantity, either by truncating $\mathbf{E}_{t-1}$ [29], or obtained with an online, local estimation algorithm [28]. These results already showed that salient transition states in the environment, such as doorways, are marked by high values of task information intake at those states. Figure 4a presents for the first time results obtained with an exhaustive sampling of trajectories traversed by an agent (see Appendix A for details), giving a much more accurate approximation of the actual quantities. For these results, and all results in the remainder of this paper, we used a policy that minimizes the relevant goal information as described in the previous section with β approaching infinity, assuring that the policy is optimal in the achieved level of performance.

These new results indicate additional secondary transition states as local maxima of the new goal information required, which lie on the crossroads between doorways, marked with × in Fig. 4a. These can be explained by the fact that the agent *needs* to make a distinction at these states between several types of goals that require distinct actions: the goals behind one doorway and those behind the other(s). As we mentioned before, making such a distinction requires goal information, and indeed, the minimum of the total relevant goal information $I(G; A_t|s_t)$ is high in these states compared to other states.

This can additionally be seen if we continue from the image of a limited working memory, and determine the amount of old information that the agent does not need anymore to select an action and can thus discard to free up its memory. This can be quantified by $I(G; \mathbf{E}_{t-1}|A_t, s_t)$, i.e. the goal information relevant to past decisions, beyond what is still relevant to the current action. This gives the picture of Fig. 4b. Here it can be seen that this quantity is also high at the doorways, resulting in a large swap of information in the working memory, whereas at the crossing points not much information can be discarded and the transition consists mostly of augmentation of information already available. This means that these are two distinct transition points that can be intrinsically distinguished via their goal information signature.

The concepts of subgoals, and task and space transitions are intuitive to humans,



(a) $p(s_t)I(G; A_t|\mathbf{E}_{t-1}, s_t)$

(b) $p(s_t)I(G; \mathbf{E}_{t-1}|A_t, s_t)$

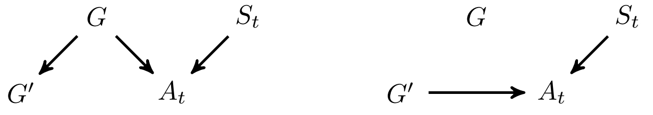
Fig. 4: Goal information transitions in a 6-room grid world navigation task. Figure (a) shows the amount of *new* goal information needed to select an action for each state, $I(G; A_t|\mathbf{E}_{t-1}, s_t)$. Figure (b) shows the amount of *old* goal information no longer needed to select an action for each state, $I(G; A_t|\mathbf{E}_{t-1}, s_t)$. Darker shades denote higher values. States that lie on the crossing points between two doorways that show local maxima in the amount of new information required are marked with × in Fig. (a).

and are readily used to describe the behavior of a biological organism or an artificial agent. Also in the creation of learning methods for artificial agents these intuitions are supplied explicitly to guide the learning process [2]. The results of this section indicate that such concepts have an immediate representation in terms of cognitive (informational) burden on the organization of the task in an agent's controller, and can be derived directly from how the task needs to be organized in the case of tight cognitive resources.

The existence of local salient (information) transitions implies that there is a natural factorization of the environment and/or taskset into separate important parts, connected by such transition points. Devising such a natural partitions in a direct and global way could be beneficial to the self-organization of an agent's behaviour, by supplying useful abstractions to guide its decision making and learning processes. In the remainder of this paper we focus on deriving such relevant, global factorizations.

5. Task Clustering

As we have stated earlier, the goal information transitions from the previous section can be explained by noting that at these transitions the agent needs to make a distinction which was not relevant at earlier steps. For instance, in our multi-room navigation scenario the agent did not have to make a distinction between the different states in the room containing its goal state before it has reached the room; it



(a) G_{in}

(b) G_{out}

Fig. 5: Bayesian networks of the goal clustering information bottleneck

could perform the same actions regardless of which of these states is its actual goal. Only after entering the room does the agent have to make a distinction between them to make sure he goes to the correct place in the room.

To put it in informational terms: in a large part of the environment the agent uses the same information to select its actions: "Which room is my goal in?" Only when it has entered the room does the goal information that is used for different goals in the same room become distinct. This suggests that one can cluster similar tasks together based on the overlap in the information about these tasks that is relevant to guiding actions. Here we will show that such a clustering can be obtained by extracting the actual relevant goal information.

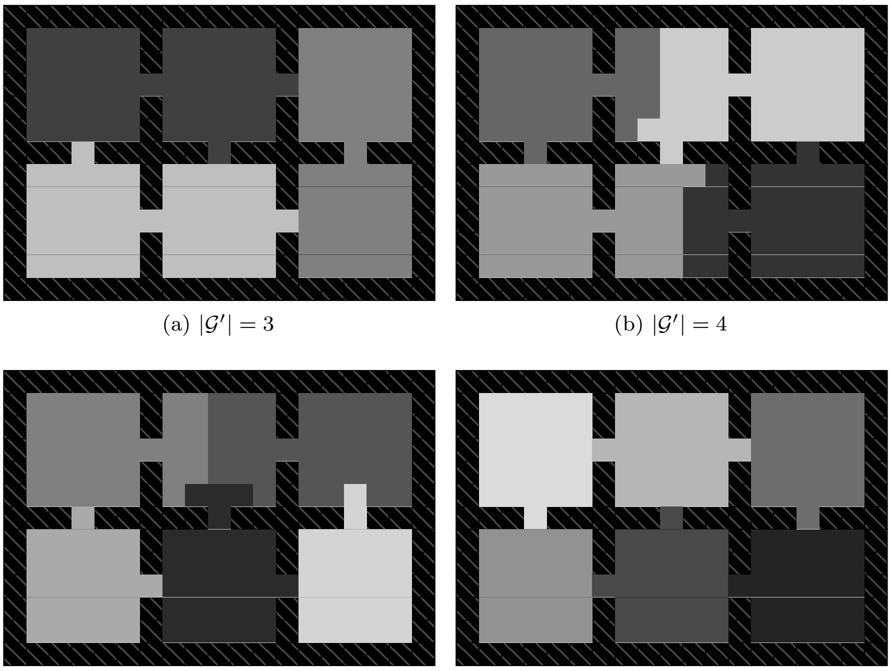
If we have obtained a policy through the RI paradigm that achieves the relevant information minimum for a given performance level, we can use the IB methods to see what the structure of this relevant information is. So, to analyze the relevant goal information, we introduce a bottleneck variable G'. The value of this variable is drawn from an alphabet $\mathcal{G}'$ of a cardinality smaller than the number of goals. We then squeeze the goal information in G that is relevant to selecting A_t through this new variable. For this problem, the networks G_{in} and G_{out} can be constructed as shown in Fig. 5, indicating that we compress G while maintaining the correct action selection given the compressed information and the state information.

Following the IB principle, our problem then becomes to minimize $I(G;G')$ over $p(g'|g)$ for fixed $I(G', S_t; A_t)$. Since $I(G', S_t; A_t) = I(G', A_t|S_t) + I(S_t; A_t)$ and $I(S_t; A_t)$ does not depend on $p(g'|g)$, we can use the simplified Lagrangian:

$$\Lambda\Big(p(g'|g), \beta\Big) = I(G;G') - \beta I(G', A_t|S_t) \tag{5}$$

We select β to be close to infinity in order to ensure maximization of $I(G', A_t|S_t)$. This results in a clustering of the goal states into a number of mutually exclusive subsets.

Figure 6 shows the results of performing this bottleneck in a 6-room navigation task for several different numbers of clusters, determined by the cardinality of the alphabet of the bottleneck variable, $|\mathcal{G}'|$. Firstly, it shows that the bottleneck is able to recover the local connectivity of the environment without an a-priori bias to do exactly this: neighboring goal cells are clustered together without an explicit concept of neighborhood in the model. Secondly, it also seems to recover global structure: the bottleneck has the tendency to cluster all goals in what we see as



(a) $|\mathcal{G}'| = 3$

(b) $|\mathcal{G}'| = 4$

(c) $|\mathcal{G}'| = 5$

(d) $|\mathcal{G}'| = 6$

Fig. 6: Goal clusters found in a 6-room grid world navigation task. Figures (a) to (d) show the mappings for increasing cardinality of the bottleneck variable. The IB in each case results in a mapping of each goal state to a single vlaue of G' with probability 1.

rooms together whenever possible. Visual inspection shows that with a cardinality of six the goal clusters coincide with the rooms especially well. While this appears to a human observer to be a very intuitive segmentation of this environment, the question arises whether this intuition about the natural organization of the task can be made objective in a way that is consistent with our informational framework.

6. How Many Task Clusters?

We want to decide what number of clusters, or what cardinality $|\mathcal{G}'|$, is the most natural to use in a given environment. This does not necessarily coincide with some notion of what the *best* amount is. For instance, if we equate 'best' with how well knowing just a goal cluster would already predict which action is chosen, compared to having the full goal information, measured by the difference between $I(G; A_t|S_t)$ and $I(G'; A_t|S_t)$, the trivial answer would be to have as many clusters as possible, with the optimum at $|\mathcal{G}'| = |\mathcal{G}|$ (note however that when data is limited this optimum can be corrected using information-theoretic concepts to help prevent

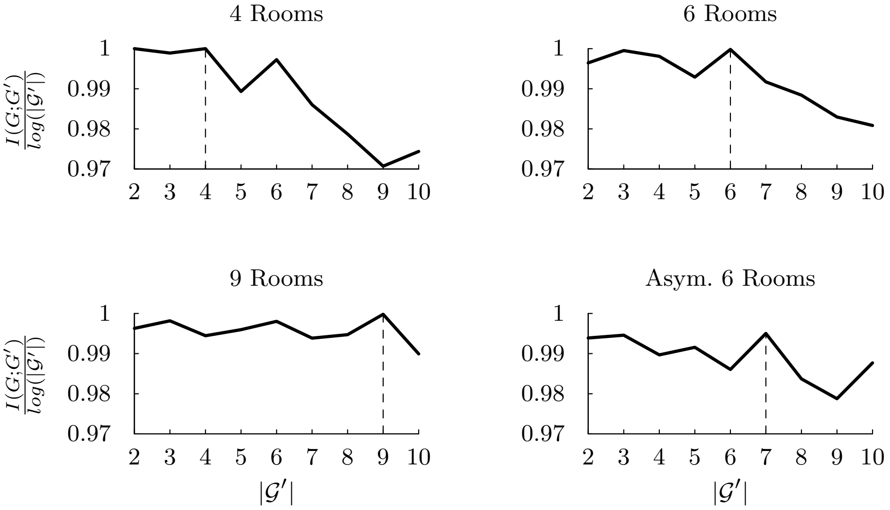


Fig. 7: Efficiency of clustering, measured by the ratio of the amount of goal information captured in the clustering, $I(G;G')$, to the total capacity of the bottleneck, $|\mathcal{G}'|$, plotted for different cardinalities of the bottleneck variable. Values closer to 1 indicate better use of the available capacity. The maximum for each scenario is marked with a dotted line.

over-fitting on sample noise [22]). Instead, we will again treat this question from the point of view of optimal use of restricted bandwidth.

Since the Markov property $G' \to G \to A_t$ holds, and G' is independant of S_t, the data processing inequality theorem [7] demands that $I(G';G) = I(G';G|S_t) \geq I(G';A_t|S_t)$. In an information bottleneck, the amount of information from the source variable(s) that is captured in the bottleneck variable(s) is often much larger than the amount of information that is retained about the target variable(s). In our scenarios for instance the latter, $I(G;G')$, turns out to be roughly five times as large as the first, $I(G';A_t|S_t)$. The compression of the source variable, not the quality of prediction, thus determines the necessary bandwidth of the bottleneck, so the effectiveness of the compression could be an important indication for the most natural choice of the number of clusters used.

The capacity of the bottleneck is limited by the cardinality of the bottleneck variable: $I(G;G') \leq \log(|\mathcal{G}'|)$. As a measurement of effectiveness we take the ratio of the actual bandwidth reached by the mapping $p(g'|g)$ to this maximum, given by $\frac{I(G;G')}{\log(|\mathcal{G}'|)}$, where a ratio approaching 1 indicates more effective usage of the total available bandwidth. The value of this ratio for a range of cardinalities and scenarios is given in Fig. 7. The scenarios used are the 2×3-room grid world of Fig. 6, the 3×3-room world of Fig. 8, a 2×2 world with the same 5×5 rooms as the previous two worlds, and the asymmetric 2×3 world of Fig. 4.

If we look at the number of clusters where the graphs have their maxima, marked by the dashed vertical lines, we see that for the first three cases, these exactly coincide with the amount of rooms in the world. In these cases, where the cardinality is equal to the number of rooms, all goal states in a single room are assigned to the same cluster, as for instance shown in Fig. 6d for the 6-room world. This clustering utilizes the bottleneck variable most effectively, getting closest to using the full available capacity, and in doing so characterizes the number of rooms that a human observer would intuitively assign to the problem.

The graphs also show other interesting factorizations as secondary and tertiary peaks. Here, the bottleneck, though not able to assign a single cluster to each room, still adheres mostly to room boundaries, such as for $|\mathcal{G}'| = 3$ in the 6-room world (cf. Fig 6a), and $|\mathcal{G}'| = 3$ and $|\mathcal{G}'| = 6$ in the 9-room world. Interestingly, for the asymmetric 6-room world, the peak does not lie at a cluster number of 6, but it turns out that a mapping with one extra cluster leads to more effectiveness, obtained by splitting the large north-eastern room into two parts that are each of a size closer to that of the other rooms.

7. Clustering Through Synergy

The factorization in the previous section captures *local* similarities in the structure of tasks, expressed in the policies that fit those tasks. This was achieved by focusing on the goal information relevant to performing those policies. In Sect. 3.3 however we showed that there is an intricate relationship between goal and state information, and that singling out only one of the two does not capture the full extent of an agent's decision making process. In this section we will focus on this *synergetic* effect, and show how this analysis exposes *global* structure in tasks and in the relationship between state and goal.

As mentioned before, the strength of the synergy between state and goal information can be judged from the difference $I(S_t; A_t|G) - I(S_t; A_t)$ (or, equally, $I(G; A_t|S_t) - I(G; A_t)$) [12, 8]. If this difference is high, the correct action can not be chosen based solely on the value of the one, but requires both to 'unlock' the available information. An example of a system with high synergy is an XOR port, where the output is 1 only when the inputs are different: knowing the value of a single input gives zero information about the output, it is *only* informative in combination with he other input.

Considering this, we will use our framework to extract the structure of the synergy between state and goal information. Similar to how in the previous section we used the IB principle to factorize the goal space to uncover the structure of the relevant goal information $I(G; A_t|S_t)$, here will use a similar method to uncover the structure in the synergy $I(S_t; A_t|G) - I(S_t; A_t)$. Again, we introduce a new variable G' that is constructed through a probabilistic mapping from G, however now we require that this variable captures not the information uniquely in G relevant to selecting actions, but instead captures the synergy. This is done by finding a mapping such that $I(S_t; A_t|G') - I(S_t; A_t)$ approaches $I(S_t; A_t|G) - I(S_t; A_t)$, and thus, considering that $I(S_t; A_t)$ is independent of the mapping, a distribution $p(g'|g)$ that maximizes $I(S_t; A_t|G')$. Such a distribution then gives a mapping with which *as much relevant state information as possible can be 'unlocked'*.

This problem can be solved in a way analogous to that of the information bottleneck, by constructing the following Lagrange equation:

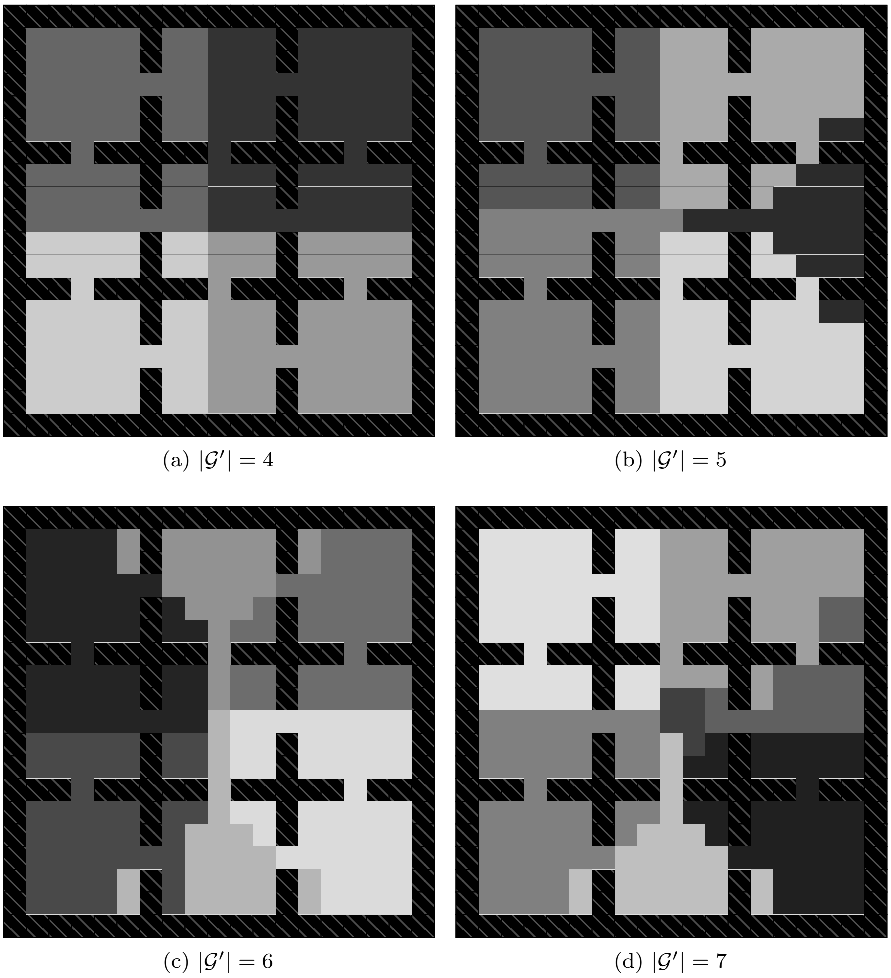
$$\Lambda\Big(p(g'|g), \beta\Big) = I(G; G') - \beta I(S_t; A_t|G'). \tag{6}$$

Note that this is not a true (multivariate) information bottleneck as presented in Sec. 3.2: it is not possible to construct a graph G_{out} that agrees with this constraint. However, because $I(S_t; A_t|G') = I(S_t, G'; A_t) - I(G'; A_t)$, we can interpret this as the information bottleneck of Fig. 5 with the extra constraint that G' should contain as little information about A_t on its own (or as a special form of an IB with side information as introduced in [6]). This again emphasizes that we are maximizing synergy.

The result of this clustering is shown in Fig. 8. As we can see, again the world is divided into several regions, however note that in this case the regions are not always constrained by walls and clusters tend to spill over into neighboring rooms. Instead of reconstructing the local connectivity of cells, these regions adhere to a more general notion of 'nearness' of cells, which transcends the walls of the world. Another interesting aspect of the patterns is that the clusters are roughly evenly distributed around the center of the environment, with small clusters marking the center of the environment when the cardinality is high enough. Thus, we see that they capture the *global* relative placing of the goal cells regardless of *local* structure created by walls, a pattern that is robust in the other scenarios presented in this paper.

The reason that this factorization appears, is because synergy arises from the importance of relative properties of states and goals. In other words, synergy appears when you need to know how a state relates to a goal in order to select an action. In our grid-world navigation examples such a property is the location of a goal state *relative* to the current state, expressed by the global direction an agent needs to travel to in order to move towards the goal. The obtained factorization captures this by forming clusters that expose the global directionality in the environment.

Our interpretation of these results is that the set of actions available to an agent (which is a part of its embodiment) can induce a geometry on the environment, in which one can formulate relationships between different states and between states and goals. These relationships are properties of the combined description of a state and a goal, which explains the high amount of synergy between the two in relation to action selection. The bandwidth constraint imposed by the bottleneck causes our clustering to capture the most global relationships, mostly disregarding local features of the environment such as walls. We speculate that this is reminiscent of how heuristics for search methods can be created by relaxing constraints on the problem of interest [13], which in a navigation task can be done by relaxing the


(a) $|\mathcal{G}'| = 4$

(b) $|\mathcal{G}'| = 5$

(c) $|\mathcal{G}'| = 6$

(d) $|\mathcal{G}'| = 7$

Fig. 8: Synergetic goal clusters found in a 9-room grid world navigation task. Figures (a) to (d) show the mappings for increasing cardinality of the bottleneck variable. The IB in each case results in a mapping of each goal state to a single vlaue of G' with probability 1.

constraints imposed by the walls and using the Manhattan distance between state and goal as a heuristic for the actual path length between them. The question of how far this parallel extends to other problems goes beyond the scope of this paper.

8. Discussion

We have laid out a unified information-theoretic framework as a tool for progressing our understanding of the structure of the task space of an agent guided by rewards, from the viewpoint of self-organization of this structure under possible constraints imposed by an agent's embrainment. Most notably, this leads to the analysis of the trade-offs between informational constraints and the performance of an agent charged with multiple tasks. We have shown that from this point of view and with the informational tool-set discussed in this paper, it is possible to derive strict bounds on an agent's information acquisition and bookkeeping capabilities, and on feasible performance when these capabilities are limited. Moreover, we showed how this framework gives a unified approach to identify a wide range of points of importance which otherwise have to be identified with manually created measures, such as salient transition points, task similarity, and both local and global structural world features. These results can shed light

We emphasize that the results presented in this paper should be also seen in a wider perspective, such as to understand how the self-organization of an agent's behaviour may be guided through task space structure and the potentiality of various goals inside this structure. Previously, we already showed that an agent can speed up learning new tasks by constructing a set of skills based on identifying salient goal information transition points, considering these as useful sub-goals and storing a sub-policy for achieving these sub-goals [28]. This is an example of so called *Transfer Learning (TL)*, currently an active topic in RL, where knowledge gained from learning a set of source tasks is transferred to a new target task [23]. More TL topics are addressed in a natural way by our framework. For instance, an important question is how to determine which source tasks to transfer knowledge from and disregard unrelated tasks, which leads to the question of how to cluster the source tasks [5]. We have shown that our framework not only directly gives a method to derive such a clustering, without the need of a seperate ad-hoc similarity measure as used in existing methods [5], but that it even gives a natural way to determine how many clusters to use. Our methods could also shed light on empirical results on the organization of cognition in biological organisms, such as recent results that show how people forget things when walking through a door [17].

Common to the results we have shown here, and to our approach in general, is that no higher level external knowledge about what may constitute useful features is used to guide the organization of an agent's behavior. For instance, we do not require a definition of transition points in explicit terms of structure of the world, such as "funnel states between highly connected areas" as commonly used in RL literature [19], or knowledge about a-priori structure designed into the MDP formulation such as some similarity measures between reward functions [5]. Our approach uncovers such concepts solely from intrinsic considerations.

We believe that the coincidence of the structure uncovered by our methods and the structural concepts employed by human intuition stems from the fact that self-organization of an agent can only be obtained by utilizing the organization of the environment and the agent's embodiment, and that this is exactly what human cognition achieves. If this hypothesis is correct, and given empirical evidence for informational constraints as a basis for the structuring of human behavior (e.g. [15]

and references therein), a closed approach to generate a family of organizational concepts in a coherent way by systematically applying cognitive constraints (in the form of information limitations) as taken in this work can constitute an important step towards guiding self-organization.

References

[1] Ay, N., Der, R., and Prokopenko, M., Information driven self-organization: The dynamical system approach to autonomous robot behavior, *Theory in Biosciences, to appear* (2011).

[2] Barto, A. G. and Mahadevan, S., Recent advances in hierarchical reinforcement learning, *Discrete Event Dynamic Systems* **13** (2003) 41–77.

[3] Bialek, W., Nemenman, I., and Tishby, N., Predictability, complexity, and learning., *Neural computation* **13** (2001) 2409–63.

[4] Blahut, R. E., Computation of channel capacity and Rate-Distortion functions, *IEEE Transactions on Information Theory* **18** (1972) 460–473.

[5] Carroll, J. L. and Seppi, K., Task similarity measures for transfer in reinforcement learning task libraries, in *The 2005 International Joint Conference on Neural Networks, (IJCNN 2005)* (2005), pp. 803–808.

[6] Chechik, G. and Tishby, N., Extracting relevant structures with side information, in *Advances in Neural Information Processing Systems 15*, eds. S Becker, S. T. and Obermayer, K., Vol. 15 (MIT Press, 2003), pp. 857–864.

[7] Cover, T. M. and Thomas, J. A., *Elements of information theory* (Wiley-Interscience, 1991).

[8] Gawne, T. J. and Richmond, B. J., How independent are the messages carried by adjacent inferior temporal cortical neurons?, *The Journal of neuroscience : the official journal of the Society for Neuroscience* **13** (1993) 2758–2771.

[9] Klyubin, A., Polani, D., and Nehaniv, C., Organization of the information flow in the perception-action loop of evolved agents, in *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, eds. Zebulum, R., Gwaltney, D., Hornby, G., Keymeulen, D., Lohn, J., and Stoica, A. (Published by the IEEE Computer Society, 2004), pp. 177–180.

[10] Klyubin, A. S., Polani, D., and Nehaniv, C. L., Keep your options open: an information-based driving principle for sensorimotor systems., *PloS one* **3** (2008) e4018.

[11] Laughlin, S. B., de Ruyter van Steveninck, R. R., and Anderson, J. C., The metabolic cost of neural information., *Nature neuroscience* **1** (1998) 36–41.

[12] McGill, W., Multivariate information transmission, *Information Theory, IRE Professional Group on* **4** (1954) 93 –111.

[13] Pearl, J., *Heuristics: intelligent search strategies for computer problem solving* (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984).

[14] Pfeifer, R., Lungarella, M., Sporns, O., and Kuniyoshi, Y., On the information-theoretic implications of embodiment – principles and methods, in *50 Years of Artificial Intelligence*, Vol. 4850 (Springer-Verlag, 2007), pp. 76–86.

[15] Polani, D., Information: currency of life?, *HFSP journal* **3** (2009) 307–16.

[16] Polani, D., Nehaniv, C., Martinetz, T., and Kim, J., Relevant information in optimized persistence vs. progeny strategies, in *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems* (Citeseer, 2006), pp. 337–343.

[17] Radvansky, G. A., Krawietz, S. A., and Tamplin, A. K., Walking through doorways

causes forgetting: Further explorations., *Quarterly journal of experimental psychology* **64** (2011) 1632–45.

[18] Salge, C. and Polani, D., Digested information as an information theoretic motivation for social interaction, *Journal of Artificial Societies and Social Simulation* **14** (2010).

[19] Şimşek, Ö. and Barto, A. G., Skill characterization based on betweenness, in *Advances in Neural Information Processing Systems 21*, eds. Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (2009), pp. 1497–1504.

[20] Slonim, N., Friedman, N., and Tishby, N., Multivariate information bottleneck., *Neural computation* **18** (2006) 1739–89.

[21] Sporns, O. and Lungarella, M., Evolving coordinated behavior by maximizing information structure, in *Artificial life X: proceedings of the tenth international conference on the simulation and synthesis of living systems* (Citeseer, 2006), pp. 323–329.

[22] Still, S. and Bialek, W., How many clusters? an information-theoretic perspective., *Neural computation* **16** (2004) 2483–506.

[23] Taylor, M. E. and Stone, P., Transfer learning for reinforcement learning domains: A survey, *The Journal of Machine Learning Research* **10** (2009) 1633–1685.

[24] Tishby, N., Pereira, F. C., and Bialek, W., The information bottleneck method, in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing* (1999), pp. 368–377.

[25] Tishby, N. and Polani, D., Information theory of decisions and actions, in *Perception-Reason-Action Cycle: Models, Algorithms and Systems*, eds. Cutsuridis, V., Hussain, A., and Taylor, J. (Springer (In Press), 2010).

[26] Touchette, H. and Lloyd, S., Information-theoretic limits of control, *Physical Review Letters* **84** (2000) 1156–1159.

[27] Touchette, H. and Lloyd, S., Information-theoretic approach to the study of control systems, *Physica A: Statistical Mechanics and its Applications* **331** (2004) 140–172.

[28] van Dijk, S. G. and Polani, D., Grounding subgoals in information transitions, in *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning* (Paris, France, 2011), pp. 105–111.

[29] van Dijk, S. G., Polani, D., and Nehaniv, C. L., 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*, eds. Fellermann, H., Drr, M., Hanczyc, M., Ladegaard, L. L., Maurer, S., Merkle, D., Monnard, P., St y, K., and Rasmussen, S. (MIT Press, Odense, Denmark, 2010), pp. 176–183.

[30] Vergassola, M., Villermaux, E., and Shraiman, B. I., 'Infotaxis' as a strategy for searching without gradients., *Nature* **445** (2007) 406–9.

Appendix A. Goal Information Transition Sampling

The following sampling method was used to find $I(G; A_t|\mathbf{E}_{t-1}, s_t)$ and obtain Fig. 4a. The mutual information between the goal and action is equal to the reduction in entropy of A_t resulting from knowing the value of G:

$$I(G; A_t|\mathbf{E}_{t-1}, S_t) = H(A_t|\mathbf{E}_{t-1}, s_t) - H(A_t|G, \mathbf{E}_{t-1}, s_t) \tag{A.1}$$

Because of the symmetry of information, it also holds that:

$$I(G; A_t|\mathbf{E}_{t-1}, s_t) = H(G|\mathbf{E}_{t-1}, s_t) - H(G|A_t, \mathbf{E}_{t-1}, s_t) \tag{A.2}$$

Considering the asymptotic equipartition theorem [7], these entropy terms can be estimated by drawing n i.i.d. samples from the combination of G, $\mathbf{E}_{t-1}$ and A_t, according to:

$$H(G|\mathbf{E}_{t-1}, s_t) \approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{g_i} p(g_i|\mathbf{e}_{t-1,i}, s_t) \tag{A.3}$$

$$H(G|A_t, \mathbf{E}_{t-1}, s_t) \approx -\frac{1}{n}\sum_{i=1}^{n}\sum_{g_i} p(g_i|a_t, \mathbf{e}_{t-1,i}, s_t) \tag{A.4}$$

$$\tag{A.5}$$

for large values of n. The probability distribution $p(g_i|\mathbf{e}_{t-1,i}, s_t)$ are obtained by applying a series of Bayesian updates $p(g_i|\mathbf{e}_k) \leftarrow \frac{1}{Z}p(a_k|s_k, g_i)p(g_i|\mathbf{e}_{k-1})$ for k from 0 to t, where Z is a normalization factor and $p(g_i|\mathbf{e}_{-1})$ is set to be uniform. Noting that $p(g_i|a_t, \mathbf{e}_{t-1,i}, s_t) = p(g_i|\mathbf{e}_{t,i}, s_t)$ gives the other distribution needed to obtain the approximation.

The estimation of $I(G; \mathbf{E}_{t-1}|A_t, s_t)$ is obtained in an analogously.

Appendix B. RI and IB Self-consistent Solutions

The following subsections derive the self-consistent solutions of the relevant information and bottle-neck type problems discussed in the text, by determining the partial derivative of the respective Lagrangian and finding the zero of that gradient.

B.1. *Relevant State Information*

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|s_t, g), \beta\Big) = I(S_t; A_t|G) - \beta E[U_G^\pi(S_t, A_t)], \tag{B.1}$$

and its partial derivative with respect to $p(a_t|s_t, g)$:

$$\frac{\partial}{\partial p(a_t|s_t, g)}\Lambda\Big(\pi(a_t|s_t, g), \beta\Big) = p(s_t, g)\log\frac{p(a_t|s_t, g)}{p(a_t|g)} - p(s_t, g)\beta U_g^\pi(s_t, a_t) \tag{B.2}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|s_t, g) = \frac{1}{Z}p(a_t|g)\exp\Big[-\beta U_g^\pi(s_t, a_t)\Big], \tag{B.3}$$

where Z is a normalization term.

B.2. *Relevant Goal Information*

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|s_t, g), \beta\Big) = I(G; A_t|S_t) - \beta E[U_G^\pi(S_t, A_t)], \tag{B.4}$$

and its partial derivative with respect to $p(a_t|s_t,g)$:

$$\frac{\partial}{\partial p(a_t|s_t,g)}\Lambda\Big(\pi(a_t|s_t,g),\beta\Big) = p(s_t,g)\log\frac{p(a_t|s_t,g)}{p(a_t|s_t)} - p(s_t,g)\beta U^{\pi}_{g}(s_t,a_t) \quad \text{(B.5)}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|s_t,g) = \frac{1}{Z}p(a_t|s_t)\exp\Big[-\beta U^{\pi}_{g}(s_t,a_t)\Big], \quad \text{(B.6)}$$

where Z is a normalization term.

B.3. *State-Goal Information Trade-Off*

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(a_t|s_t,g,\beta)\Big) = \alpha I(S_t;A_t|G) + (1-\alpha)I(G;A_t|S_t) - \beta E[U^{\pi}_{G}(S_t,A_t)], \quad \text{(B.7)}$$

and its partial derivative with respect to $p(a_t|s_t,g)$:

$$\begin{aligned}\frac{\partial}{\partial p(a_t|s_t,g)}\Lambda\Big(\pi(a_t|s_t,g),\beta\Big) &= \alpha p(s_t,g)\log\frac{p(a_t|s_t,g)}{p(a_t|s_t)}+ \\ &\quad (1-\alpha)p(s_t,g)\log\frac{p(a_t|s_t,g)}{p(a_t|s_t)}- \\ &\quad p(s_t,g)\beta U^{\pi}_{g}(s_t,a_t) \quad \text{(B.8)}\end{aligned}$$

Equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(a_t|s_t,g) = \frac{1}{Z}p(a_t|s_t)^{\alpha}p(a_t|g)^{1-\alpha}\exp\Big[-\beta U^{\pi}_{g}(s_t,a_t)\Big], \quad \text{(B.9)}$$

where Z is a normalization term.

B.4. *Goal Clustering*

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(g'|g),\beta\Big) = I(G;G') - \beta I(G';A_t|S_t). \quad \text{(B.10)}$$

Taking its partial derivative with respect to $p(g'|g)$, equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(g'|g) = \frac{1}{Z}p(g)\exp\Big[\sum_{s_t}p(s_t)D_{KL}\Big(p(a_t|g,s_t)||p(a_t|g',s_t)\Big)\Big], \quad \text{(B.11)}$$

where Z is a normalization term, and the Kullbeck-Leibler divergence $D_{KL}\Big(p(a_t|g,s_t)||p(a_t|g',s_t)\Big) = \sum_{a_t}p(a_t|g,s_t)\log\frac{p(a_t|g,s_t)}{p(a_t|g',s_t)}$.

B.5. *Synergetic Clustering*

Here, the Langrangian is given as:

$$\Lambda\Big(\pi(g'|g), \beta\Big) = I(G; G') - \beta I(S_t; A_t|G') \tag{B.12}$$

$$= I(G; G') - \beta\Big(I(S_t, G'; A_t) - I(G', A_t)\Big). \tag{B.13}$$

Taking its partial derivative with respect to $p(g'|g)$, equating this derivative to zero and rearrangement of terms leads to the self-consistent solution:

$$p(g'|g) = \frac{1}{Z} p(g) \exp\Big[\sum_{s_t} p(s_t) D_{KL}\Big(p(a_t|g, s_t)||p(a_t|g', s_t)\Big) - D_{KL}\Big(p(a_t|g)||p(a_t|g')\Big)\Big], \tag{B.14}$$

where Z is a normalization term, and the Kullbeck-Leibler divergences $D_{KL}\Big(p(a_t|g, s_t)||p(a_t|g', s_t)\Big) = \sum_{a_t} p(a_t|g, s_t) \log \frac{p(a_t|g,s_t)}{p(a_t|g',s_t)}$ and $D_{KL}\Big(p(a_t|g)||p(a_t|g')\Big) = \sum_{a_t} p(a_t|g) \log \frac{p(a_t|g)}{p(a_t|g')}$.