Goal-Directed Empowerment: Combining Intrinsic Motivation and Task-Oriented Behaviour

Nicola Catenacci Volpi[1], Daniel Polani[1]
[1]School of Engineering and Computer Science, University of Hertfordshire, Hatfield, AL109AB, UK

Empowerment is an information-theoretic measure representing the capacity of an agent to affect its environment. It quantifies its ability to inject information in the environment via its actions and to recapture this information through its sensors. In a nutshell, it measures the number of future options available and perceivable by the agent. Originally, the definition of empowerment does not depend on any particular extrinsic goal and it is determined only by the interaction of the agent with the world and the structure of its action-perception cycle. In this paper we introduce a new formalism that combines empowerment maximization with externally specifiable goal-directed behaviour. This has two main implications: on the one hand, the study of the relationship between empowerment optimization and goal-directedness, to investigate to which extent these two desirable behaviours can co-exist; on the other hand, from a more operational point of view, the derivation of a method to generate a behaviour (i.e., a policy of a Markov decision process) that is both empowered and goal-directed, in order to design agents capable of being as "empowered" as possible when facing any extrinsic task. Finally, we study how this hybrid policy is able to handle problems of uncertain or changing goals and delayed goal commitment.

Index Terms—**Intrinsic Motivation, Empowerment, Goal-directed Behaviour, Information Theory, Robustness**

I. Introduction

EMPOWERMENT [1], [2] is an intrinsic motivation measure that quantifies the amount of control an agent has over its environment and its capacity to perceive such control through its sensors. It is defined information-theoretically as the Shannon channel capacity [3] of the actuation channel of the action-perception loop of an agent. Intuitively, it quantifies in bits the amount of controllable options that are also observable and available to an agent in the future. Although a pure empowerment-maximizing agent is often not strictly optimal in terms of the optimization of explicit external reward functions, behaviour based solely on empowerment maximization sometimes coincides with the behaviour induced by "natural" reward functions such as, for example, in the cases of a pendulum, a bike, or a marine vehicle being balanced in their upright poses [4], [5]. In addition, empowerment maximization has been proved to be beneficial in several diverse scenarios from Artificial Intelligence (AI) [1], [6]–[12] to robotics and control [1], [4], [5], [13]–[15]. Salge et al. [2] have hypothesized that natural organisms tend to maximise their empowerment and that this behaviour is in general beneficial and plausible for both natural and artificial agents. Still, during their lifetime, organisms are forced to perform activities that are imposed through the external constraints that survival imposes on them and that cannot be expressed through empowerment maximization, but require explicit goals to be formulated. In addition, in some applications artificial agents may maximise their empowerment to aid the achievement of a given extrinsic goal in order to tackle the task in a more robust and controllable manner or to enhance exploration when learning is involved. Hence, it was suggested that empowerment optimisation should be performed in conjunction with other tasks [12], [14].

In this study, we modify empowerment maximization in a way that permits us to express a new quantity that agents can optimise when they want to be as empowered as possible within the limits imposed by another active goal. We have named this quantity *goal-directed empowerment* (GDE). The introduction of this new concept will allow us to answer fundamental questions regarding the relationship between empowerment and goal-directed behaviour: to which extent agents can be empowered when they are also constrained by doing other activities and how the empowerment of an agent decreases when a particular goal is chosen. In other words, we address whether there is a way to measure the change in the empowerment landscape of an agent when its behaviour becomes goal-directed. A further aim is to investigate the problem of loss of expected return (i.e., the regret) incurred to allow the agent to maintain a desired level of empowerment during the solving of a Markov decision process (MDP) [16]. The GDE formalism introduces a new quantitative approach tackling these issues. We will see that a commitment to a specific goal may need to be paid for by loss of empowerment. This poses the question of why agents should sacrifice their rewards to keep their empowerment large. Although to be empowered usually implies to lose reward, its real advantages have to be found in respect of long term gains in highly dynamic scenarios: for instance, where the goals of agents change often and unexpectedly. In these situations empowerment and goal-directed behaviour are not only alternative to each other, but they can also work together to satisfy both known short-term needs of the agent and unknown long-term ones. Therefore, we propose that agents that are pursuing a goal can benefit from keeping a certain amount of empowerment available for the sake of robustness. GDE is a generalisation of the empowerment formalism whereby agents can benefit from being empowered in all those tasks where this

This work was partially supported by EC Horizon 2020 programme under the project WiMUST (Grant agreement no: 645141, Strategic objective: H2020 - ICT-23-2014 - Robotics). Corresponding author: Nicola Catenacci Volpi (email: n.catenacci-volpi@herts.ac.uk).

is possible. Hence, GDE considerably widens the applicability of empowerment, allowing roboticists and scientists from the artificial intelligence community to design agents that are both empowered and goal-directed in all those tasks that are not compatible with the sole maximization of classical empowerment. While the present study is focused on agents with discrete actions and simple grid-world environments with discrete state space, in the Conclusion (Section IV) we will discuss how the presented formalism could be extended to address more complex scenarios (i.e., continuous domains with partial observability).

The rest of the paper will be structured as follows: in Section I-A we will examine the literature relevant to our study; in Section II we will present the goal-directed empowerment formalism; in Section III a set of simulated experiments will be reported and their results discussed; finally, in Section IV we will provide our conclusion, including a perspective on future work.

A. Related Work

1) Empowerment as Intrinsic Motivation

In recent years, instead of considering models of explicit tasks to be learned by artificial agents, increasingly researchers have begun to consider models of *intrinsic motivation* [17], [18]. These are behaviour-generating models which do not rely on an external reward function, but instead try to shape the behaviour according to some plausible principles which aim to be either biologically or psychologically plausible, or else attempt to exploit universal aspects of the sensorimotor interaction of the agent with the world. The study of intrinsic motivations in natural organisms aims at identifying the underlying principles with the possibility to implement them in robots and artificial agents in order to produce universal behaviour. When the goal of an agent is induced by the satisfaction of intrinsic motivations, the optimization of these measures results in the autonomous generation of behaviour. For instance, homeokinesis [19] and predictive information [20] have been optimised to produce robot behaviour without relying on any external reward function, taking into account the robot embodiment and the nature of its interactions with the world only [21]. Work that applies intrinsic motivations to machine learning usually uses these measures as incentives to improve the agent's learning speed and quality, acting as guidance to the underlying exploration process [22]–[24]. In this regard, there has been interest in combining intrinsic motivation measures with MDPs, in particular to enhance reinforcement learning (RL) exploration and model acquisition [25]–[29]. What distinguishes these approaches from ours is their focus on learning and exploration rather than on robustness and viability of goal-directed behaviour.

Empowerment [1], [2] is an intrinsic motivation measure that can generate behaviour in a task-independent and universal manner. Its formalism uses information theory to model the joint concepts of having an agent in control of its own environment and its ability to perceive this controllability through its sensors. Empowerment measures the maximum amount of information that an agent can transmit to its future sensory perception from its present actuators through the environment (we will formally define empowerment in Section II-A2). It quantifies (in bits) the capacity of an agent to influence its own future. States with a large numbers of potential distinct futures are states with large empowerment values. On the contrary, if the agent has no impact on the environment that it can sense, the agent attains only the minimum value of empowerment, which is 0 bits. When there is no specific goal to pursue, it is desirable to be in states where actions have the largest effect on the environment and to keep as many available options as possible in sight of an unknown future. This concept of preparedness motivates the introduction of empowerment as information-theoretic quantification and formalization of the amount of open options that an agent can both control and perceive. The principle of empowerment maximisation has been applied to very different areas of artificial intelligence [1], [6], [7], [30], [31], robotics and control [1], [4], [5], [13]. Importantly, empowerment was defined in the same fashion across all these experiments. In addition, it has been shown that when an agent maximises its empowerment, it aims at its self-preservation (since both death or complete breakdown implies zero empowerment [7], [15]). Hence, empowerment can be used as an early warning metric of an agent reaching the borders of its viability domain.

2) Empowerment, Goals and Reinforcement Learning

Although empowerment maximization on its own is capable of solving certain AI problems, it is obviously not the best method to use for all of them, because the state with maximum empowerment will not always correspond with our expectation of a goal state in general, even if it does so in unexpectedly many cases. Here, instead of considering exclusively tasks where the goal is in the state with maximum empowerment, we would like to target arbitrary goals, whilst retaining as many benefits of an empowered strategy as possible. Starting from the work of Mohamed and Rezende [8] , which uses variational inference in deep neural network to offer a computationally efficient approximation of empowerment, several approaches have been proposed to combine empowerment with RL (e.g., see [11], [32]). Although all these studies combine empowerment with MDPs, their main scope is the improvement of the agent's learning of a model of the environment, aiding the exploration process underlying model-based RL. In contrast, in this paper we integrate empowerment with MDP planning and pure reward maximisation with no learning involved. In a more similar vein with this paper, the empowered skills method [14] uses a Lagrangian approach in the direct policy search framework to maximise a trade-off between reward and empowerment. This method uses an approximation of empowerment (i.e., entropy of future outcomes) that is exact for deterministic environments. The result of this optimisation is a policy that exhibits a largely diverse behaviour, which have been shown to be advantageous in adversarial scenarios such as table-tennis. Leibfried et al. [12] also use a Lagrangian-like approach to compute a policy that maximises a trade-off between reward and empowerment. They provide a Bellman-like optimality principle for the corresponding optimisation process and the method has been shown to improve the agent performance at the initial stages of interaction with the environment using

an action-critic algorithm for highly dimensional robotic tasks (e.g., ant and humanoid simulated robots). In this paper, we also present a way to compute a policy that agents can use to maximises a trade-off between empowerment and reward, showing also how this behaves in tasks with varying goals (see Section III-D). However, it is important to underline that the main scope of our study is not to improve the performance of RL, but to introduce the new information-theoretic quantity *goal-directed empowerment*. This fuses the advantages of the empowerment formalism with the requirements of behaviour directed towards a given goal, enabling to investigate the change of empowerment landscape when an agent has goals to pursue. We also expect that using GDE agents could behave in a more robust and controllable manner.

II. The Goal-directed Empowerment formalism

A. Technical Preliminaries

1) Markov Decision Processes

Markov Decision Processes [16] are sequential models of decision-making used to represent the behaviour of agents that act in stochastic environments with the aim of maximising their expected return. Formally, an undiscounted MDP is defined as the tuple $(\mathcal{S}, \mathcal{A}, P, R)$, where this is composed as follows: $\mathcal{S}$ is the set of states where the agent can be; $\mathcal{A}$ is the set of actions the agent can choose in every state; $P(s'|a, s)$[1] represents the probability that the agent reaches the state $s' \in \mathcal{S}$ after having executed the action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$; $R(s, a, s')$, with reward function $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, is the reward that the agent receives when, after having executed the action a being in state s, arrives in state s'. Within the MDP framework, a deterministic *policy* $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping between states and actions that is used to represent possible behaviours of an agent, defining for every state which action to take. An optimal policy, denoted as π^*, is a policy that from every state guarantees to the agent the maximum expected return. Where, denoting with s_t and a_t the state and the action taken by the agent at time step t, the *return* is defined by $G \doteq \sum_{t=1}^{\infty} R(s_t, a_t, s_{t+1})$. In other words, we have $\pi^*(s_t) = \arg\max_\pi \mathbb{E}[\sum_{t=1}^{\infty} R(s_t, \pi(s_t), s_{t+1})]$. Given a policy π, the expected return cumulated starting in state s is called the *state-value* function of s for the policy π and is denoted by $V^\pi(s)$. The *action-value* function $Q^\pi(s, a)$ is the expected return cumulated starting in state s taking action a first and then following π. An optimal policy π^* induces the optimal state-value and action-value functions, which are denoted by $V^*(s)$ and $Q^*(s, a)$ respectively (and which are, unlike π^*, unique).

2) The Empowerment formalism

Given a horizon h, the h-step empowerment is defined as the capacity of the *h-step actuation channel* of an agent. The time horizon h represents the number of steps ahead in the future that are considered in evaluating the empowerment of the agent. Let us denote by $a^h \doteq a_1 \cdot a_2 \cdot a_3 \cdot \cdots \cdot a_h \in \mathcal{A}^h$ an action sequence of length h, being $\mathcal{A}^h$ the set of all possible action sequences of length h. Given that the agent is in state s at time t, the h-step actuation channel $(A_t^h, P(S_{t+h}|A_t^h, S_t = s), S_{t+h})$ is a communication channel (see Section 2 of the Supplementary Material). Its source is the random variable A_t^h, representing the possible action sequences of length h beginning in state s at time t and executed in an open-loop fashion. Its receiver is the random variable S_{t+h}, which is the state reached by the agent after h steps. Empowerment is defined in an open-loop manner, meaning that the probing action sequences a^h do not utilize any feedback during their execution. To complete the definition of the channel it is necessary to introduce the h-step transition probabilities $P(s'|a^h, s)$, which can be derived using simple algebraic steps (see [4]). This conditional distribution plays the role of the actuation channel, with the h-step actuations being its input and the resulting state its output. Note that only the effects of the agent's actions on the environment within the time horizon h are captured by empowerment. After having defined the h-step actuation channel in state s, we can now define the *h-step empowerment* of state s as its Shannon capacity as follows

$$\mathfrak{E}^h(s) \doteq \max_{P(a^h|s)} I(S_{t+h}; A_t^h | S_t = s) \tag{1}$$

where $I(X; Y)$ denotes the mutual information between the random variables X and Y (see Section 2 of the Supplementary Material). Being a Shannon channel capacity, empowerment is an information-theoretic quantity measured in bits. The computation of $\mathfrak{E}^h(s)$ can be done using the standard Blahut-Arimoto (BA) algorithm [2], [33]. Note that usually empowerment is defined as the Shannon capacity of the channel from the agent's actuators to its sensors. However, in the case where the agent has full observability over the state space, the sensor variable and the state variable can be equated without loss of generality. Then, empowerment becomes a measure of the influence that an agent has on the whole environmental state space. In this paper we will always assume full observability.

B. Goal-directed Empowerment: definition and computation

1) GDE Definition

We here introduce goal-directed empowerment (GDE), which is a generalization of empowerment for goal-oriented behaviour. As in traditional MDP models, goal-directedness is modelled by the optimization of expected return and the performance of extrinsic tasks is represented by the execution of a given policy π. Within this framework, we define goal-directed empowerment as follows:

> *The h-step GDE of state s is the Shannon capacity of an actuation channel that considers only those action sequence distributions that guarantee on average a desired level of expected action-value in s.*

We formulate this as a constrained channel, as depicted in Figure 1. We name it *goal-directed actuation channel*. The desired minimum level of future average performance is defined by a chosen lower bound $\bar{Q}_s$ of $\mathbb{E}[Q(s, A_t^h)]$, which is

[1] In this paper we will use the shorthand notation $P(x)$ to denote $Pr(X = x)$.

the action-value function in s averaged over the optimising distributions of all the possible future action sequences of length h. In this paper we will always assume Q being the optimal action-value function Q^*. However, GDE could also be parametrised by a suboptimal Q function in order to handle situations where a suboptimal action-value might be more appropriate, for example in the case of a less informed agent. The h-step action-value function $Q(s, a^h)$ of an action sequence $a^h \doteq a_1 \cdot a_2 \cdot a_3 \cdot \cdots \cdot a_h$ is defined as the return averaged over all the possible outcomes of the action sequence plus the action-value $Q(s_{h+1}, a_{h+1})$. Hence, $Q(s, a^h) \doteq \mathbb{E}[G^h + Q(s_{h+1}, a_{h+1})|a_1, \ldots, a_h, s]$, where $G^h \doteq \sum_{t=1}^{h} R(S_t, a_t, S_{t+1})$ is the return random variable with $S_1 = s$. Hence, given an action-value function parameter $Q(s, a^h)$, a threshold $\bar{Q}_s$ and assuming that the agent is in state s at time t, the *h-step goal-directed empowerment* $\bar{\mathfrak{E}}^h_{\bar{Q}}(s; Q)$ is defined as the solution of the following constrained optimization problem:

$$\begin{aligned} \bar{\mathfrak{E}}^h_{\bar{Q}}(s; Q) \doteq \max_{P(a^h|s)} \quad & I(S_{t+h}; A^h_t | S_t = s) \\ \text{s. t.} \quad & \mathbb{E}[Q(s, A^h_t)] \geq \bar{Q}_s \end{aligned} \tag{2}$$

Note that the h-step GDE is parametrised by the h-step

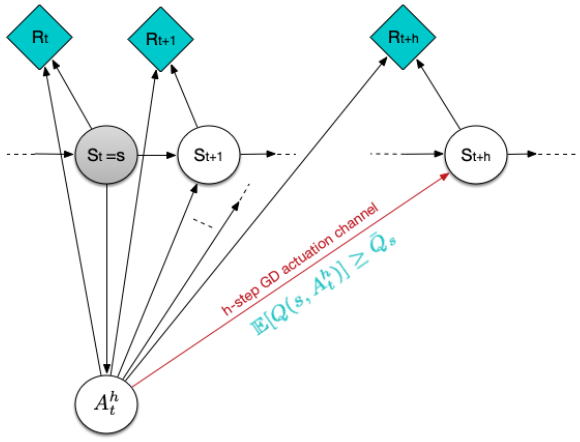


Fig. 1: Dynamic Bayesian network representing the MDP of an agent that that at time t is in state s (in grey) and that operates with action sequences. The *h-step goal-directed actuation channel* is highlighted in red.

action-value function $Q(s, a^h)$, which is used to ignore all the distributions of the action sequences starting in s that do not guarantee at least an average action-value $\bar{Q}_s$.[2] Since to execute h-step action sequences that maximise the action-value $Q(s, a^h)$ implies the accomplishment of the task modelled by the given MDP, to impose a threshold $\bar{Q}_s$ over the average action-value will enforce the empowerment optimisation to take into account only action sequences with a certain level of goal-directedness. These actions sequences will be closer to optimality as the chosen threshold will be larger and thus will tend to be more goal-directed. Here the action-value of a state is computed by averaging over all available action sequences, but the threshold $\bar{Q}_s$ should be different in each state. This choice was made because setting a global action-value threshold (i.e., one averaged over both the state and action sequence distributions) would not specify how the action distributions are selected for each individual state. There would be no obvious choice for an ad hoc attribution of action-value per state with a global threshold, which is exacerbated by the fact that for different states s empowerment differs in general and is moreover achieved by different action sequence distributions in each state. We will tackle the problem of how to choose the $\bar{Q}_s$ in Section III. Table I lists the mathematical notation used throughout the paper to facilitate the reader in following the subsequent material.

2) GDE Computation: GDE-BA algorithm

The h-step GDE can be computed using the *GDE-Blahut–Arimoto (GDE-BA) algorithm* described below. The h-step action-value function $Q(s, a^h)$ is one of its inputs. Hence, before running the procedure a method like the Value Iteration algorithm [16], [25] (reported in Section 1 of the Supplementary Material) has to be used to compute $Q(s, a)$. Then, the h-step action-value function parameter $Q(s, a^h)$ can be computed using the aforementioned definition. Once the h-step action-value function has been computed and the desired $\bar{Q}_s$ has been chosen for all s, the computation of $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$ follows a BA procedure similar to the one of classical empowerment. The main difference is represented by the presence of a term composed of $Q(s, a^h)$ and a Karush–Kuhn–Tucker (KKT) multiplier μ_s. The latter is a parameter that can be used to tune the amount of goal-directedness in the GDE of state s. In the following, we will present the GDE-BA algorithm to compute the GDE as a function of μ_s. How μ_s is linked to $\bar{Q}_s$ will be discussed in the next section. To compute GDE, the GDE-BA algorithm uses an iterative procedure that produces consecutive estimates $P_{\bar{\mathfrak{E}}}(a^h|s)_k$ of the GDE-maximizing action sequence distribution $P_{\bar{\mathfrak{E}}}(a^h|s)$ (i.e., the $\arg\max$ of Equation (2))[3], where k denotes the iteration step, and estimates $I_{\bar{\mathfrak{E}}}(S_{t+h}; A^h_t = a^h, S_t = s)_k$ of the true conditional mutual information $I_{\bar{\mathfrak{E}}}(S_{t+h}; A^h_t = a^h, S_t = s)$ (Equation (3)). In this regard, for each iteration k, $I_{\bar{\mathfrak{E}}}(S_{t+h}; A^h_t = a^h, S_t = s)_k$ can be computed as follows (for more details about the derivation of Equation (3) see [4]):

$$I_{\bar{\mathfrak{E}}}(S_{t+h}; A^h_t = a^h, S_t = s)_k = \sum_{s_{t+h} \in \mathcal{S}} P(s_{t+h}|a^h, s) \log \left(\frac{P(s_{t+h}|a^h, s)}{\sum_{b^h \in \mathcal{A}^h} P(s_{t+h}|b^h, s)) P_{\bar{\mathfrak{E}}}(b^h)_k} \right) \tag{3}$$

Before starting the iterative procedure, the probability distribution $P_{\bar{\mathfrak{E}}}(a^h)_k$ has to be initialized for $k = 0$. This can be done, for instance, by starting with a uniform distribution as estimate, hence $\forall a^h \in \mathcal{A}^h \; P_{\bar{\mathfrak{E}}}(a^h)_0 \doteq \frac{1}{|\mathcal{A}^h|}$. Then, for each iteration $k = 1, 2, \ldots$ the following recursive equations are

[2] In the following, for the sake of notational convenience, to denote GDE we will use $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$ instead of $\bar{\mathfrak{E}}^h_{\bar{Q}}(s; Q)$, having the action-value function parameter implicitly assumed.

[3] Although the GDE-maximizing action sequence distribution $P_{\bar{\mathfrak{E}}}$ is always conditioned on the state s, to have a more compact notation, in the rest of the paper we will often use the shorthand notation $P_{\bar{\mathfrak{E}}}(a^h)$ to denote $P_{\bar{\mathfrak{E}}}(a^h|s)$.

TABLE I: GLOSSARY of MATHEMATICAL SYMBOLS

Symbols	Meaning
h	time horizon
a^h	h-step action sequence
$P(a^h\|s)$ or $P(a^h)$	probability of action sequence a^h in state s (s sometime may be implicitly assumed)
$\mathfrak{E}^h(s)$	classical h-step empowerment of state s
$P_{\mathfrak{E}}$	empowerment-maximizing probability distribution of action sequences
$Q(s,a^h)$	MDP action-value of action sequence a^h in state s
$\bar{Q}_s$	GDE action-value threshold in state s
$\bar{\mathfrak{E}}^h_{\bar{Q}}(s;Q)$ or $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$	h-step GDE of state s with action-value threshold $\bar{Q}$ and function Q (Q sometime may be implicitly assumed)
μ_s	trade-off KKT parameter, used to tune goal-directedness vs empowerment
$P_{\bar{\mathfrak{E}}}$	GDE-maximizing probability distribution of action sequences
$\pi_{\bar{\mathfrak{E}}}$	GDE policy
$Q_{\bar{\mathfrak{E}}}(s,a)$	GDE action-value of action a in state s
$\bar{Q}^h_{\bar{\mathfrak{E}}}(s)$	action-value of state s averaged over the h-step GDE-maximizing action sequences' probability distribution

used to get the new distribution $P_{\bar{\mathfrak{E}}}(a^h)_k$ from the previous one $P_{\bar{\mathfrak{E}}}(a^h)_{k-1}$, up to convergence towards the final estimate of the GDE-maximizing action sequence distribution $P_{\bar{\mathfrak{E}}}(a^h)$:

$$\forall\, a^h \in \mathcal{A}^h \quad P_{\bar{\mathfrak{E}}}(a^h)_k = \\ = \frac{1}{\bar{Z}_k} P_{\bar{\mathfrak{E}}}(a^h)_{k-1} e^{I_{\bar{\mathfrak{E}}}(S_{t+h};A^h_t=a^h,S_t=s)_{k-1}+\mu_s Q(s,a^h)} \tag{4}$$

The normalization factor $\bar{Z}_k$ is given by

$$\bar{Z}_k \doteq \sum_{a^h\in\mathcal{A}^h} P_{\bar{\mathfrak{E}}}(a^h)_{k-1} e^{I_{\bar{\mathfrak{E}}}(S_{t+h};A^h_t=a^h,S_t=s)_{k-1}+\mu_s Q(s,a^h)} \tag{5}$$

Using $I(S_{t+h};A^h_t|S_t=s) = \sum_{a^h\in\mathcal{A}^h} P(a^h|s) I(S_{t+h};A^h_t = a^h|S_t=s)$, the GDE estimate at iteration step k can be computed with the following equation

$$\bar{\mathfrak{E}}^h_{\bar{Q}}(s)_k = \sum_{a^h\in\mathcal{A}^h} P_{\bar{\mathfrak{E}}}(a^h)_k I_{\bar{\mathfrak{E}}}(S_{t+h};A^h_t=a^h,S_t=s)_k \tag{6}$$

The algorithm is iterated until the absolute value of the difference of two consecutive estimates of GDE is below a very small threshold ϵ, in other words, when $\mid \bar{\mathfrak{E}}^h_{\bar{Q}}(s)_k - \bar{\mathfrak{E}}^h_{\bar{Q}}(s)_{k-1} \mid \leq \epsilon$. When ϵ is small enough, we can use the estimate for GDE as an approximation for the real GDE value $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$.

3) GDE-BA Algorithm Derivation

Here we give the highlights of the mathematical derivation of the GDE-BA algorithm. In Section 3 of the Supplementary Material we report the complete proof. The constrained optimisation problem (2) that finds the GDE-maximizing action sequence distribution $P_{\bar{\mathfrak{E}}}(a^h)$ at a given level of action-value $\bar{Q}_s$ can be turned into an unconstrained one using the Lagrangian method with KKT conditions. These are necessary because the optimization problem contains the utility inequality constraint. First, let us define the auxiliary function $\phi(s'|s,a^h) \doteq \frac{P(s'|s,a^h)}{\sum_{b^h\in\mathcal{A}^h} P(b^h)P(s'|s,b^h)}$. Then, using Equation (3) we can write

$$I(S_{t+h};A^h_t|S_t=s) = H(A^h) + \\ + \sum_{a^h\in\mathcal{A}^h}\sum_{s_{t+h}\in\mathcal{S}} P(s_{t+h}|s,a^h)P(a^h)\log\big(\phi(s_{t+h}|s,a^h)P(a^h)\big) \tag{7}$$

The Lagrangian $\mathcal{L}$ with KKT conditions of the optimization problem (2) is

$$\mathcal{L} \doteq I(S_{t+h};A^h_t|S_t=s) + \lambda\left(1 - \sum_{a^h\in\mathcal{A}^h} P(a^h)\right) \\ + \mu_s\left(\sum_{a^h\in\mathcal{A}^h} P(a^h)Q(s,a^h) - \bar{Q}_s\right) \tag{8}$$

where the Lagrangian multiplier λ ensures the normalization of the probability distribution of action sequences and the KKT multiplier μ_s is chosen for the action sequence distribution to fulfil the utility constraint. In addition, we will see that the found solution for $P_{\bar{\mathfrak{E}}}(a^h)$ will be always non-negative. We can use the Lagrangian condition $\forall\, \tilde{a}^h \in \mathcal{A}^h \;\; \frac{\partial\mathcal{L}}{\partial P(\tilde{a}^h)} = 0$ to solve the maximization problem (2) for $P(\tilde{a}^h)$ as follows

$$\frac{\partial\mathcal{L}}{\partial P(\tilde{a}^h)} = 0 \Rightarrow P(\tilde{a}^h) = \\ = e^{\sum_{s_{t+h}\in\mathcal{S}} P(s_{t+h}|s,\tilde{a}^h)\log_2\left(P(\tilde{a}^h)\phi(s_{t+h}|s,\tilde{a}^h)\right) - \lambda + \mu_s Q(s,\tilde{a}^h)} \tag{9}$$

The multiplier λ is eliminated by normalization, leading to the normalizer $\bar{Z}$ of Equation (5) and to Equation (4) for $P_{\bar{\mathfrak{E}}}(\tilde{a}^h)$. There, we have obtained $I_{\bar{\mathfrak{E}}}(S_{t+h};A^h_t=\tilde{a}^h,S_t=s)$ using Equation (3) and the definition of $\phi(s'|s,a^h)$. Looking at Equation (4), we can see that the resulting $P_{\bar{\mathfrak{E}}}(\tilde{a}^h)$ is always non-negative. Finally, to relate the parameter μ_s with the threshold $\bar{Q}_s$, given the KKT conditions $\mu_s \geq 0$ and $\mu_s\left(\bar{Q}_s - \sum_{a^h\in\mathcal{A}^h} P(a^h)Q(s,a^h)\right) = 0$, we have to distinguish the case when $\mu_s = 0$ from the one with $\mu_s > 0$. For $\mu_s = 0$, the $P_{\bar{\mathfrak{E}}}(\tilde{a}^h)$ given by Equation (4) is equal to the empowerment-maximizing action sequence distribution

$P_{\mathfrak{E}}(a^h)$. The latter, using the BA algorithm for the computation of classical empowerment [2], is given by

$$P_{\mathfrak{E}}(a^h) = \frac{1}{Z} P_{\mathfrak{E}}(a^h) \exp(I(S_{t+h}; A_t^h = a^h, S_t = s)) \quad (10)$$

where Z is used for normalization. Indeed, for $\mu_s = 0$, we can see that Equation (4) and Equation (10) are identical. Hence, when the empowerment-maximizing action sequence distribution allows the agent to gather an amount of expected action-value that is larger or equal than $\bar{Q}_s$ we can set $\mu_s = 0$. For $\mu_s > 0$, we plug Equation (4) into the KKT condition related with the utility and solve numerically to get the value of μ_s that corresponds to the desired threshold $\bar{Q}_s$. If the resulting equation has no solution, then there is no solution that satisfies the utility constraint with the given action-value function Q and threshold $\bar{Q}_s$.

C. The Goal-directed Empowerment Policy

Within the context of classical empowerment agents self-generate their own behaviour ascending locally the gradient of empowerment, which sooner or later will bring them to the state with locally maximal empowerment. Also in the case of GDE, since $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$ quantifies the number of options with average action-value larger or equal than $\bar{Q}_s$ available in states s, an agent that follows its gradient will be placed in the state with most sufficient high value options. Hence, to follow the GDE of an agent does not necessarily imply to increase its expected return, because, according to GDE, what counts is the number of options with large value and not the value itself. Therefore, walking along the GDE gradient allows the agent to be as much empowered as possible fulfilling the utility constraint but not necessarily to behave in a goal-directed manner (e.g., to reach a desired goal state). This is the reason why, to generate a behaviour that is both empowered and goal-directed, in this section we introduce the GDE policy $\pi_{\bar{\mathfrak{E}}}$. In this regard, the classical approach to derive an optimal policy in stochastic sequential decision-making would be to use a Bellmann equation that combines immediate reward with the future return. However, in the context of GDE, we are not only interested in future return but in its combination with empowerment, which is the projected GDE for the future. This will be larger if more of the achievable values in the next step are higher. So, the GDE policy should choose a successor state that has more high-value potential future states as successors (or else, fewer, but very high value ones, as per nature of the Lagrangian).

The aforementioned observations led to the following definition of GDE policy $\pi_{\bar{\mathfrak{E}}}$ that allows agents to take actions that optimise a GDE action-value function $Q_{\bar{\mathfrak{E}}}(s,a)$, which combines $\bar{\mathfrak{E}}^h_{\bar{Q}}(s)$ and $Q(s,a^h)$ at the given trade-off parameter μ_s. This GDE policy $\pi_{\bar{\mathfrak{E}}}$ can be used to make an agent address the task given by the optimization of the action-value function parameter, being at the same time as much empowered as possible given the desired levels of performance μ_s. The procedure to compute $\pi_{\bar{\mathfrak{E}}}$ is inspired by the Value Iteration algorithm. Assuming that $I_{\bar{\mathfrak{E}}}(S_{t+h}; A_t^h = a^h, S_t = s')$ for KKT parameter μ_s and $Q(s,a^h)$ have already been computed, then the computation of $Q_{\bar{\mathfrak{E}}}(s,a)$ and $\pi_{\bar{\mathfrak{E}}}(s)$ proceeds as reported in Table II. Note that the $P_{\bar{\mathfrak{E}}}(a^h|s')$ is a GDE-achieving distribution *but not* the actual policy that the agent is following - that policy is given by $\pi_{\bar{\mathfrak{E}}}$. Furthermore, differently from the Value Iteration algorithm, $Q_{\bar{\mathfrak{E}}}(s,a)$ is computed using only one iteration, because the action-value function of the left-hand side of Equation (11) is not as the one on the right-hand side, therefore there is no bootstrapping or recursion. This decision was made on purpose. In an initial study, a double iteration algorithm was implemented to guarantee self-consistency between GDE and action-value (i.e., to compute the $Q^k_{\bar{\mathfrak{E}}}$ induced by $\pi^k_{\bar{\mathfrak{E}}}$ in iteration k, then to obtain a new $I^{k+1}_{\bar{\mathfrak{E}}}$ with that $Q^k_{\bar{\mathfrak{E}}}$ as parameter together with the novel corresponding $\pi^{k+1}_{\bar{\mathfrak{E}}}$, afterwards obtaining from this policy $\pi^{k+1}_{\bar{\mathfrak{E}}}$ a new $Q^{k+1}_{\bar{\mathfrak{E}}}$ in the next iteration $k+1$ and so forth), but the double iteration converged to an oscillating solution; in other words, no self-consistent formulation of $Q_{\bar{\mathfrak{E}}}$ could be constructed. Hence, the method implemented by Equations (11) and (12) has been adopted, which decouples the value-function Q that represents the goal-directedness of the agent from the value-function $Q_{\bar{\mathfrak{E}}}$ that is used to generate a GDE behaviour. This approach is exact for an optimal action-value parameter Q^* and can be considered an approximation for arbitrary action-value functions.

III. Experiments

In the following experiments we will consider episodic MDPs with the following structure: each state $s \in \mathcal{S}$ indicates a different location in the world. The action set $\mathcal{A} = \{\uparrow, \rightarrow, \downarrow, \leftarrow, \cdot\}$ contains the possible one-step movements that the agent can do towards the directions north, east, south, west plus the "stay" action, which makes the agent stay where it is. Regarding the transition probabilities $P(s'|s,a)$, if not stated differently, an action succeeds with probability 0.8 or moves the agent perpendicularly with probability 0.1 for each direction (never backwards). If the agent tries to move into a wall or towards the edge of the world, it remains where it is. The reward function R is defined as follows: in every state the agent receives reward -1 except for the goal state where the agent receives reward 0. In particular, moving into a wall is not punished specially, but incurs just the cost due to the lost time. The optimal h-step action-value function $Q^*(s,a^h)$ will be used as GDE parameter. Usually, Lagrangian are based on some kind of universal threshold but, since in GDE every state s has its own threshold $\bar{Q}_s$, here we face the problem of how to choose these $\bar{Q}_s$ for each state. Due to the highly inhomogeneous nature of the system's dynamics, a global rule that would allow the $\bar{Q}_s$ to adapt for each state (such as a rule of the form $\bar{Q}_s = \mathbb{E}[Q^*(s, A_t^h)] - \Delta Q$ with constant ΔQ parameter), would result in local adjustments that would be not easily comparable with each other. Therefore, we instead use the same value μ of the Lagrangian parameters μ_s of all states, in other words for all $s \in \mathcal{S}$ we select $\mu_s = \mu$. Note that having the same μ_s for every state does not necessarily imply having the same $\bar{Q}_s$ for every state, because μ_s represents the trade-off between both action-value and empowerment and is not exclusively linked to $\bar{Q}_s$.

TABLE II: GDE Policy Computation

$$\forall\, s \in \mathcal{S}, a \in \mathcal{A} \quad Q_{\bar{\mathfrak{E}}}(s,a) = \sum_{s' \in \mathcal{S}} P(s'|s,a) \left(R(s',s,a) + \sum_{a^h \in \mathcal{A}^h} P_{\bar{\mathfrak{E}}}(a^h|s') \left(\mu_s Q(s',a^h) + I_{\bar{\mathfrak{E}}}(S_{t+h}; A_t^h = a^h, S_t = s') \right) \right) \tag{11}$$

$$\forall\, s \in \mathcal{S} \quad \pi_{\bar{\mathfrak{E}}}(s) = \arg\max_a Q_{\bar{\mathfrak{E}}}(s,a) \tag{12}$$

A. Trading-off Empowered and Optimal Policies

Since GDE is a formalism for empowered and goal-directed decision-making combined, in this section we will show that moving μ to its extreme values, i.e. 0 or infinity, $\pi_{\bar{\mathfrak{E}}}$ takes the form of a pure empowerment-maximizing policy in the first case and of the classical optimal MDP solution in the second. In practical terms, a large μ will be enough to approximate μ going to infinity. In addition, we will look at what kind of $\pi_{\bar{\mathfrak{E}}}$ one obtains with values of μ that lie between these extremes.

Consider the 9x9 grid world reported in Figures 2.a-c, where wall cells are coloured in white and the absorbing goal state is labelled with the letter G (later on we will consider non-absorbing goals as well). In addition, the red arrows in each cell represent the actions chosen by $\pi_{\bar{\mathfrak{E}}}$ in the corresponding states with the dot indicating "stay". Each state is coloured according to the two-step $\bar{\mathfrak{E}}^2_\mu$ of that state (see the associated colour bars for the corresponding values in bits).[4] In Figure 2.a, with $\mu = 0$, for every state $\bar{\mathfrak{E}}^2_\mu$ is equal to the 2-step classical empowerment $\mathfrak{E}^2$, whose action sequence distribution $P_{\bar{\mathfrak{E}}}(a^2)$ achieves a state averaged action value of $\bar{Q}^2_{\bar{\mathfrak{E}}} = -8.4$, where $\bar{Q}^h_{\bar{\mathfrak{E}}}(s) \doteq \mathbb{E}_{\bar{\mathfrak{E}}}[Q(s, A^h_t)] = \sum_{a^h \in \mathcal{A}^h} P_{\bar{\mathfrak{E}}}(a^h|s) Q(s,a^h)$. Furthermore, the behaviour obtained by executing the resulting $\pi_{\bar{\mathfrak{E}}}$ is identical to classical empowerment maximization obtained through local gradient ascent. In Figure 2.c, $\bar{\mathfrak{E}}^2_\mu$ and $\pi_{\bar{\mathfrak{E}}}$ are reported for $\mu = 128$. From Equation (4) we know that large values of μ imply strong goal-directedness, which causes the action-value to become $\bar{Q}^2_{\bar{\mathfrak{E}}} = -6.63$ and $\bar{\mathfrak{E}}^2_\mu$ to drop to 0 bits in most states, as it is possible to see from the figure (note that this MDP has few degenerate solutions due to its transitions' noise and topology). This happens every time there are only few possible ways to meet the selected utility constraint. In contrast to that, the four states with $\bar{\mathfrak{E}}^2_\mu \simeq 0.4$ bits (in light blue) have multiple 2-step action sequences that guarantee an average action-value larger than $\bar{Q}_s$. The GDE policy of Figure 2.c shows that the chosen value for μ is large enough to make $\pi_{\bar{\mathfrak{E}}}$ equal to π^*. In addition to the extreme values of μ, other intermediate $\pi_{\bar{\mathfrak{E}}}$ can be computed using in-between μ values, which allows the agent to trade-off empowerment maximization against a quick arrival to the goal. In this regard, in Figure 2.b we reported $\pi_{\bar{\mathfrak{E}}}$ for $\mu = 0.25$, where $P_{\bar{\mathfrak{E}}}$ achieves $\bar{Q}^2_{\bar{\mathfrak{E}}} = -8.08$. In this case the agent always arrives to the goal passing, when possible, through states with large empowerment.

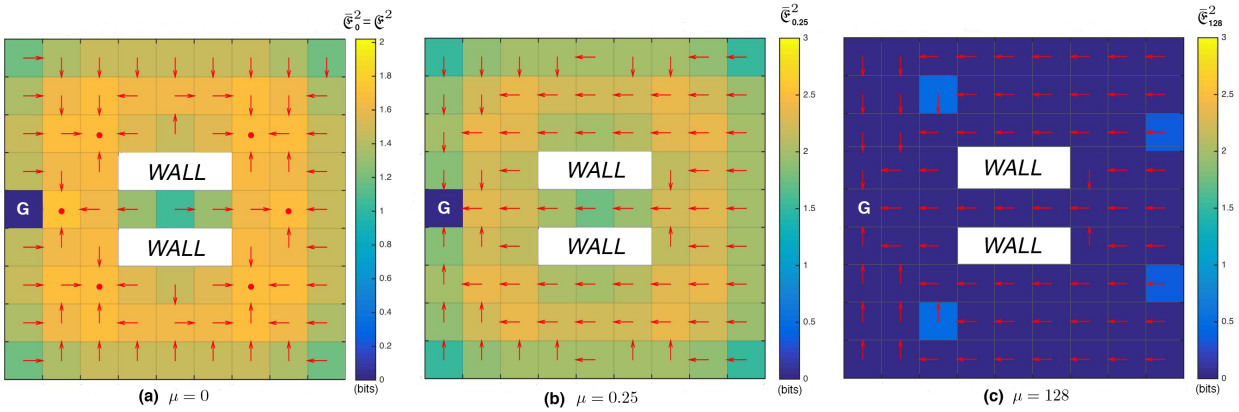
[4]Since in the following experiments we will investigate the GDE as a function of μ_s and for every μ_s there a corresponding $\bar{Q}_s$ (see Section II.B.3), in the following we will write $\bar{\mathfrak{E}}^h_{\mu_s}(s)$ instead of $\bar{\mathfrak{E}}^h_{\bar{Q}_s}(s)$ to denote GDE.

B. To what extent can agents be empowered in reaching a goal?

In this section we investigate the extent an agent can be empowered when it has a goal to reach. We will see that in general the GDE of an agent decreases when the agent has a task to face, because goal-directedness prunes the spectrum of options of the agent towards those that are more favourable towards the completion of the given task. This phenomenon expresses itself differently in different states, for instance according to their distance from the goal.

First, let us analyse the "opportunity cost" relationship between GDE and regret, showing how state averaged $\bar{\mathfrak{E}}^h_\mu(s)$ depends on state averaged regret ρ^h, where $\rho^h(s) \doteq V^*(s) - \bar{Q}^h_{\bar{\mathfrak{E}}}(s)$. In Figure 3, for different values of μ, we consider $\bar{\mathfrak{E}}^2_\mu$ as a function of ρ^2, both averaged over an uniform distribution of states, for the grid world considered in the previous section. The plot shows that the average $\bar{\mathfrak{E}}^2_\mu$ is a monotonically increasing function of ρ^2. For $\mu = 0$, $\rho^2 = 1.78$ and the average $\bar{\mathfrak{E}}^2_0$ has a value of 1.93 bits, which is the full average empowerment available to the agent prior to any decision about pursuing a goal. As μ becomes large, both $\bar{\mathfrak{E}}^2_\mu$ and ρ^2 decrease to 0. For intermediate values of μ the trade-off curve shows the average empowerment left to the agent when it behaves at the chosen level of performance. This amount of available empowerment can be explained by the fact that in many states the agent can choose two perpendicular actions to arrive to the goal that are similar in terms of obtained action-value. In general, different MDPs will have different $\bar{\mathfrak{E}}^h_\mu / \rho^h$ trade-off curves, which can be used to characterise tasks in terms of amount of empowerment needed to perform the task at the desired level of performance and the amount of empowerment still available when the agent is achieving this performance. The difference between the average $\bar{\mathfrak{E}}^h_\mu$ left at trade-off parameter μ and average classical empowerment $\mathfrak{E}^h$ can be interpreted as the amount of empowerment invested by the agent to attain at least the target average action-value.

In Figure 4, $\bar{\mathfrak{E}}^2_1$ and $\pi_{\bar{\mathfrak{E}}}$ are shown in a larger maze with a more complicated topology. Note the three straight sequences of states coloured in orange in the middle of the three rooms of the environment. Along these routes toward the goal, $\bar{\mathfrak{E}}^2_1$ is lower than their surrounding states. This happens because in those states the GDE-maximizing action sequence distribution $P_{\bar{\mathfrak{E}}}$ is squeezed towards those action sequences that allow the agent to take the shortest straight route to the goal and all other sequences would make the agent lose return, violating the utility constraint associated with the given μ. Hence, those corridors represent the lack of options in those states. On the



(a) $\mu = 0$ **(b)** $\mu = 0.25$ **(c)** $\mu = 128$

Fig. 2: **(a-e)** 9x9 grid world environments. Goal states are marked with the letter "G". The red symbols within each cell represent the action taken by $\pi_{\bar{\mathfrak{E}}}$ in that state, with red circles representing stay actions. The cells are coloured according to $\bar{\mathfrak{E}}^2_\mu$, whose values in bits are indicated in the reported color bars. **(a)** For $\mu = 0$ ($\bar{Q}^2_{\bar{\mathfrak{E}}} = -8.4$) the $\bar{\mathfrak{E}}^2_0$ landscape is equivalent to the one of 2-step classical empowerment $\mathfrak{E}^2$ and $\pi_{\bar{\mathfrak{E}}}$ coincides with the classical empowerment maximizing behaviour. **(b)** With $\mu = 0.25$ ($\bar{Q}^2_{\bar{\mathfrak{E}}} = -8.08$) $\pi_{\bar{\mathfrak{E}}}$ is significantly more goal-directed. **(c)** For $\mu = 128$ ($\bar{Q}^2_{\bar{\mathfrak{E}}} = -6.63$) $\pi_{\bar{\mathfrak{E}}}$ is equivalent to the optimal policy π^* and $\bar{\mathfrak{E}}^2_{128}$ drops close to 0 bits for almost all the states.

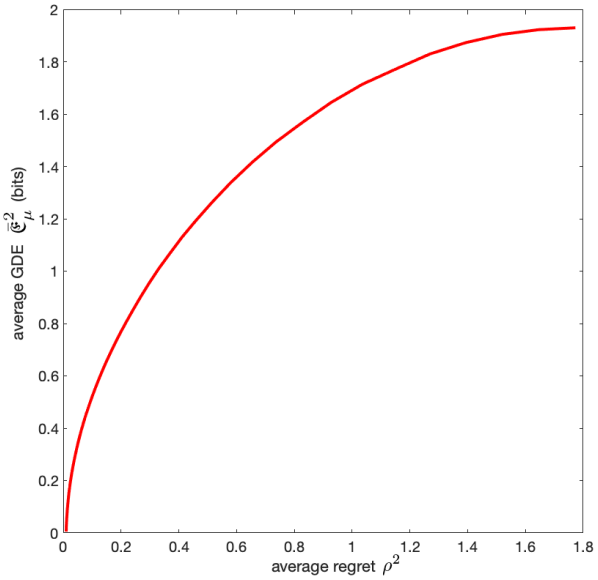
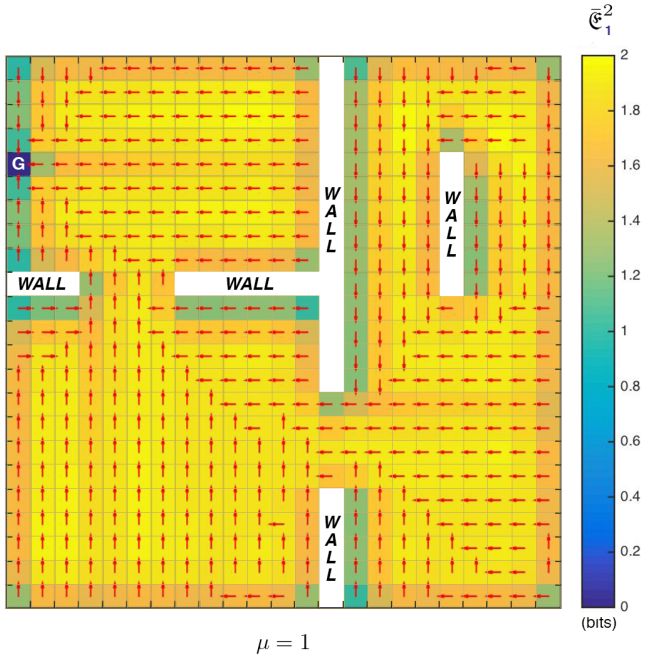


Fig. 3: Trade-off curve for varying μ of the regret ρ^2 versus 2-step GDE for the MDP presented in the previous section. Both quantities are averaged across all the states using an uniform distribution.

contrary, the other yellow areas contain states where several action sequences give a similar outcome in terms of Q^*, so these sequences provide to the agent more available options to fulfil the utility constraint and consequently they increase $\bar{\mathfrak{E}}^2_1$. Within these orange paths, the drop of $\bar{\mathfrak{E}}^2_1$ is more significant as the agent gets closer to the goal. This happens because when the desired performance is not stringent (i.e., low μ), while the agent is far from the goal it has several alternative routes to reach it. By contrast, as the agent approaches the goal, the number of ways to reach it decreases so that the agent needs to become more precise regarding what he does. Then, in states adjacent to the walls the limitation of the movement of the agent is reflected by a decrease of $\bar{\mathfrak{E}}^2_1$, as for classical empowerment. In the case of GDE, $\bar{\mathfrak{E}}^2_1$ drops more for the states adjacent to the walls that are located on the side away from the goal (coloured in green), because in these states also the option of moving away from the goal has a low probability in $P_{\bar{\mathfrak{E}}}$. We conclude that to reach a goal makes an agent lose



$\mu = 1$

Fig. 4: 23x23 grid world environment. We can observe low GDE corridors given by GDE-maximizing action sequences that are more probable than others when the agent is obliged to meet the desired level of performance. In states along these corridors $\bar{\mathfrak{E}}^2_1$ decreases more for states that are closer to the goal.

its h-step empowerment when there are only few different h-step action sequences that initiate the routes towards the goal at the desired level of utility. Hence, in states where the agent

is forced to be more precise to meet the given performance constraint, $\bar{\mathfrak{E}}^h_\mu$ is usually small.

C. "Stop, it is too risky there..."

We would attribute "caution" to agents that do not like to be confined inside cluttered spaces, because once they enter there is no a quick way to go out. Sometime they stop because they are too afraid of the unexpected dangers that could appear in front of them, or because they are waiting for better opportunities somewhere else. So, they can decide not to move if to proceed could impact their freedom to find a new route, which would allow agents to react to sudden events on time. Similarly, a GDE-maximizing agent may decide to stop at the borders of areas of the environment where empowerment is low, refusing to further increase its return, in scenarios where the goal is adjacent to walls or obstacles. In Figure 5, we report a grid world where the goal is at the edge of the environment and different 4-step GDE policies are depicted for increasing values of μ. In this scenario the agent decides to stop before arriving too close to the goal due to its proximity to the edge. In Figures 5.a,b, μ is increased from 0.03 ($\bar{Q}^4_{\bar{\mathfrak{E}}} = -14.61$) to 0.04 ($\bar{Q}^4_{\bar{\mathfrak{E}}} = -14.49$), causing the agent to stop increasingly closer to the goal as it becomes more goal-directed. Finally, in Figure 5.c we observe that with a value of 0.17 ($\bar{Q}^4_{\bar{\mathfrak{E}}} = -13.88$), μ is large enough to make the agent reach the goal from every state of the environment. A similar stopping behaviour of that shown in Figures 5.a,b can be found every time the agent needs to pass through states with low empowerment to arrive to the goal. Indeed, if to take any alternative routes to the goal implies also a loss of return (i.e., longer routes), the agent can decide to stop where it is. For instance, consider an agent situated in front of the entrance of a narrow tunnel, which must be traversed whilst following the shortest route to the goal (as depicted in Figure 5). Consider also that all the alternative routes to the goal are considerably longer than the shortest one. Should the agent go through the tunnel or should it take the way around it? On the one hand, to go through the tunnel implies to decrease its empowerment (inside the tunnel there are very limited options). On the other hand, to reach the goal going around the tunnel increases the length of the route to the goal. Hence, when $\mu = 0.03$, to not lose the overall $\bar{\mathfrak{E}}^4_{0.03}$ the agent does not commit to either of these two alternatives but rather stops at the entrance of the tunnel (see Figure 5.a). For a slightly larger trade-off parameter ($\mu = 0.04$) the agent decides to enter the tunnel, but it still does not proceed completely, stopping one state after the entrance (Figure 5.b). Finally, for $\mu = 0.17$, the increased goal-directedness causes the agent to traverse the tunnel and to reach the goal (Figure 5.c).

In general, to be in the most empowered state is beneficial for agents when the goal is completely unknown and its location will be revealed at later stages, because probably some of the options left available maximising empowerment will be useful when the goal will be known. Although previous work showed that global measures such as graph centrality are strongly related with empowerment, which is a local quantity, empowerment is not equivalent to pure centrality or reachability measures (for a comparison of these concepts see [5], [6]), which could be used to get similar effects (e.g., using a Laplacian diffusion). Namely, when the interaction of the agent with the environment is stochastic, the options provided by empowerment maximization are controllable options - as opposed to centrality, which does not cater for random noise. The Laplacian looks at diffusion - but does not distinguish whether this has been generated by the environment or by the agent. The stopping behaviour generated for small μ by GDE maximization could be interpreted as the one of an agent that does not fully commit to the goal represented by the Q^* parameter nor to the possibility of a new forthcoming goal that could unexpectedly substitute the previous one and placed in any state of the environment. Hence, the agent decides to stay in a position that in the future could serve both goals. Furthermore, in Figure 5 a larger μ implies the agent stopping closer to the given goal, as if the agent believes more that goal to be the real one as the trade-off parameter increases. So, although at first glance to favour empowerment over return could seem not convenient for a greedy agent, this is not the case in the long term when multi-task scenarios are considered. Research on MDPs with uncertain goals has been conducted by the AI community in the last years [34], [35]. In reward-uncertain MDPs the reward is defined as a set representing known bounds, or other imprecise parameters about the reward, and criteria as min-max regret are used to obtain robust solutions for the worst case scenario. Although the scope of GDE is not to solve optimally problems with uncertain goals, which are presented here for the sake of interpreting GDE, it is worth to mention that its maximization make agents satisfy other desirable properties under an unique framework, such as increased states' reachability, agent's self-preservation and controllability (i.e., noise adversity).

D. Robustness over Changing Goals

The GDE policy has the ability to make an agent arrive to its goal through highly empowered routes. This usually happens paying some cost in terms of return. In this section we aim to address the following questions: why should an agent choose to be empowered at the price of losing reward? More precisely: why should the agent stop when to proceed further towards the given goal implies decreasing its empowerment? In the following experiment we will investigate the robustness of $\pi_{\bar{\mathfrak{E}}}$ in a task where the goal is suddenly moved to a new random location and compare its performance with that of π^*. Consider the grid world reported in Figure 6, where the agent is located in the state indicated with the letter S and has to arrive to the goal state labelled with the letter G. As for the task presented in the previous section, the optimal policy π^* takes the shortest route to the goal, which traverses a tunnel where the states have low empowerment. With $\mu = 0.03$, the 3-step GDE policy makes the agent take an alternative route from S to G, which is longer and more empowered (see the green path in Figure 6.a). This goes around the tunnel instead of going through it or stopping in front of its entrance. To evaluate the robustness of the two policies over an unexpected change of goal, we compared their performances

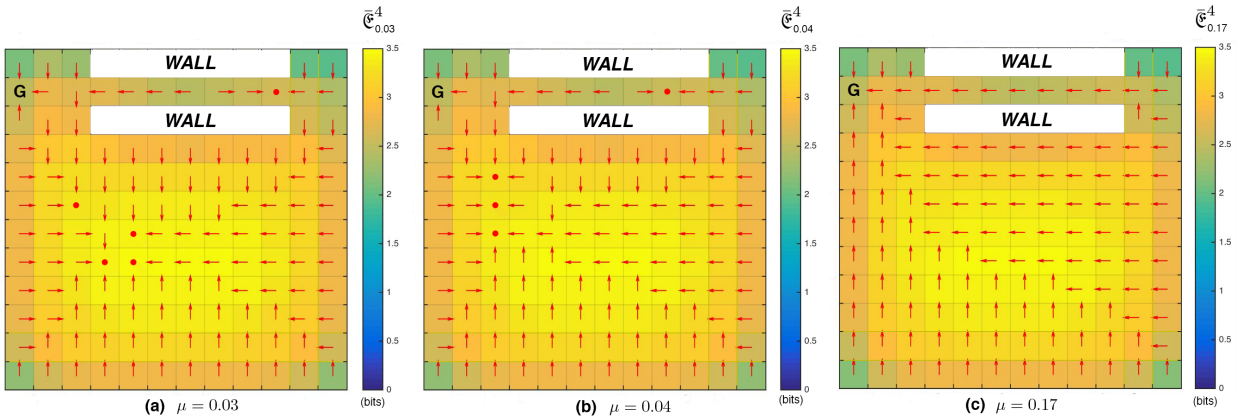


Fig. 5: 12x12 grid world environments with a tunnel. Cells are coloured according to $\bar{\mathfrak{E}}^4_\mu$. μ is increased from a minimum of 0.03 in (a), to 0.04 in (b) and 0.17 in (c).

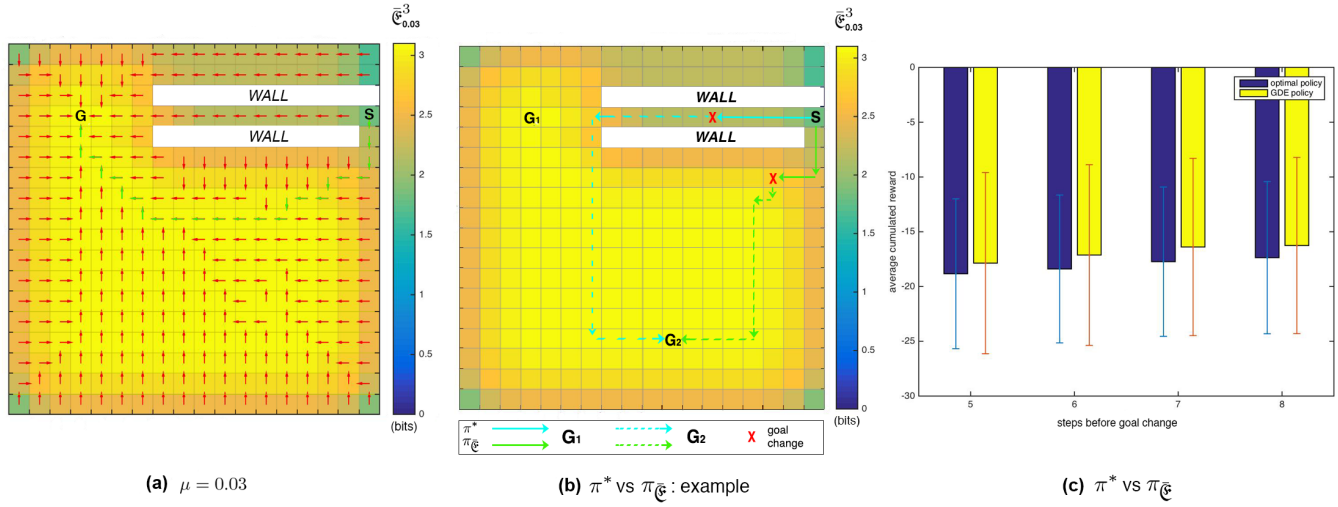


Fig. 6: (a-c) 18x18 grid world environments. **(a)** For $\mu = 0.03$ $\pi_{\bar{\mathfrak{E}}}$ makes the agent take an empowered route that goes around the tunnel (in green). **(b)** Example trial where both $\pi_{\bar{\mathfrak{E}}}$ (azure) and π^* (green) attempt to bring the agent to the first goal G_1 (solid arrows). Here, after five steps the goal is changed to a new random location G_2 (where the red cross is depicted). Then, the optimal and GDE agents proceed to G_2 using new $\pi_{\bar{\mathfrak{E}}}$ and π^* respectively (dashed arrows). **(c)** The average returns of $\pi_{\bar{\mathfrak{E}}}$ and π^* against the number of steps after which the goal is changed. The returns are averaged across 10000 simulations where the locations G_2 is chosen randomly for each trial.

(i.e., average return) when, after they started seeking a first goal, the goal position randomly changes according to an uniform distribution after n steps from the beginning of the trial. The agent has no prior knowledge about when this change will happen and where the new goal will be. If the agent starts one trial using π^* for the first goal, it will use a new π^* for the new goal. Similarly, if the agent starts a trial following $\pi_{\bar{\mathfrak{E}}}$, it will use a new $\pi_{\bar{\mathfrak{E}}}$ to reach the new goal. In Figure 6.b, we report an example of this two-phased task, showing the execution of the two policies for a particular instance. Figure 6.c compares the average return over 10000 simulations for $n = 5, 6, 7, 8$, cumulated starting when the goal is changed until the agent reaches the new goal. As it is possible to see from the bar plots, when the agent visits more empowered states using $\pi_{\bar{\mathfrak{E}}}$, it does not perform worse than π^* and it might be slightly better.

E. Noise Adversity and Delayed Commitment

One of the main properties of empowerment is that it decreases when there is a source of noise in the environment. Indeed, an empowerment-maximizing agent does not only prefer to have as many available options as possible, but it also wants to have control over the outcomes of its choices. This property is retained by the GDE formalism as well. In the following experiment we will show that between two goals, a more empowered $\pi_{\bar{\mathfrak{E}}}$ favours the one that is not surrounded by noisy states. Furthermore, along the resulting route, the choice between the two goals is delayed by $\pi_{\bar{\mathfrak{E}}}$ as much as the utility constraint allows it (similarly to the "delayed choice" phenomenon described in [36]), showing a lack of commitment between them until this becomes strictly necessary. In the grid world represented in Figure 7 we have an agent starting in the state marked with the letter S. The two goal states are indicated by G_1 and G_2 and have reward 0. The agent receives a reward of -1.5 when it lands in the

penalty cells, which are marked with the letter P. Entering every other state has reward -1. Regarding the transition model, when the agent executes one of its actions, it succeeds with probability 0.99 or it slips to each one of the two perpendicular directions with probability 0.005. In addition, when the agent is in the states indicated by the letter N the probability of slipping is 0.2, making these states noisier than the rest of the environment. In short, to arrive at G_1 the agent must pass through penalty states, however to arrive at G_2 requires passing through noisy states. Therefore, the agent has to choose whether to go through the penalty states incurring loss of reward, or to go through the noisy states incurring loss of control. We can get both of these behaviours within the GDE framework by tuning the trade-off parameter μ. With $\mu = 10.2$ ($\bar{Q}^4_{\bar{\mathfrak{E}}} = -4.84$), when the agent follows the 4-step GDE policy (azure trajectory in Figure 7), it passes through the noisy states and reaches G_2, showing the same behaviour of an optimal policy that maximizes the return averaged over the noise. With $\mu = 6.7$ ($\bar{Q}^4_{\bar{\mathfrak{E}}} = -6.88$), $\pi_{\bar{\mathfrak{E}}}$ passes through the a penalty state to arrive to G_1 (red trajectory), showing that the agent now accepts the loss of reward in order to avoid the loss of empowerment it would incur by approaching the noisy states. The two trajectories reported in Figure 7 show another

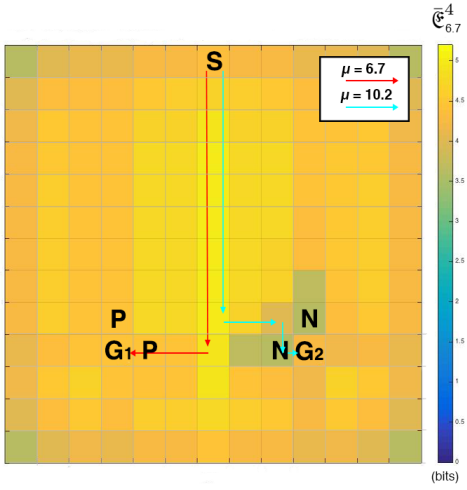


Fig. 7: 13x13 grid world environment. The agent starts in the state marked with the letter S and it can choose to reach one of the two alternative goals G_1 and G_2. There are two penalty states, marked with the letter P, which have a negative reward of -1.5 instead of reward -1 that all the other states have, apart from the goals where the reward is 0. The state marked with the letter N are noisy states. When the agent tries to move to a noisy state there is a probability of 0.2 to slip to one of the two perpendicular states, instead of slipping with probability 0.01 as for every other state. The paths taken by two agent that follow the two GDE policies, generated using $\bar{\mathfrak{E}}^4_{\mu}$, are depicted by arrows, where the red ones represent the route taken using $\pi_{\bar{\mathfrak{E}}}$ for $\mu = 6.7$ and the azure ones represent the route taken when $\mu = 10.2$.

interesting behaviour which occurs when a GDE-maximizing agent must choose between more than one goal. The color map overlaid on the grid world of Figure 7 represents $\bar{\mathfrak{E}}^4_{6.7}$. In the middle of the environment there is a vertical "highway" of states with large GDE (coloured with the brightest yellow) which both $\pi_{\bar{\mathfrak{E}}}$ cause the agent to traverse. These states have $\bar{\mathfrak{E}}^4_{6.7}$ and $\bar{\mathfrak{E}}^4_{10.2}$ large because most of their $P_{\bar{\mathfrak{E}}}$ are allocated to action sequences that move the agent towards both G_1 and G_2. The two considered policies pass through these states up to the locations where the decision regarding which goal to reach must necessarily be made. We see a delayed commitment between these behaviours that leaves the options about which goal to choose open for as long as possible.

IV. Conclusion

In this paper we introduced goal-directed empowerment, a new information-theoretic quantity that generalises the empowerment formalism to the case when agents have to tackle an externally specifiable task. Whenever possible, GDE allows agents to retain the benefits of aiming to be as empowered as possible for all the tasks where the goal is not in the state with maximum empowerment. The GDE framework allowed us to investigate the relationship between empowerment and goal-directed behaviour, i.e. to study the change of empowerment landscape when an agent has a given goal to pursue. Generally, the presence of a goal causes the empowerment of an agent to decrease, because it prunes the full spectrum of available options of the agent leaving only those that are necessary to achieve the goal. By analysing the GDE of an agent, it is possible to quantify the amount of empowerment lost by the agent to achieve its goal at the chosen level of performance and the amount of empowerment still available when the agent is tackling the goal. In this study we have also devised the GDE policy $\pi_{\bar{\mathfrak{E}}}$, which can be used by agents to both solve a task and be empowered for different trade-off levels μ. Interestingly, for low values of μ, agents are not always capable of arriving to their goal and sometime they decide to stop before reaching it, which poses the following question: why agents should pay in terms of return to keep their empowerment? One way to answer this question is that agents keep their empowerment to lose it at later stages when new unknown forthcoming tasks could be assigned to them.

In robotics, empowerment can be used to characterise safe and reliable Human-Robot Interaction (HRI) [15]. For application purposes, since a robot companion is usually expected to provide user-centric services to humans, a method that would allow it to deliver a service while complying with the requirements specified by empowerment-based HRI would be highly desirable. The GDE formalism could be used to achieve this using a principled mathematical framework. Although in this paper GDE was applied to simple grid world experiments and its formalism was presented for fully observable systems with discrete state and action spaces, its information-theoretic formulation within the MDP framework is universal enough to allow its extension towards more complex scenarios. Partial observability is a built-in feature of the empowerment formalism. Thus, the challenge of incorporating partial observability in GDE lies on the MDP side of the framework. In this regard, initial information-theoretic treatment of passive POMDPs has been described in [37], which could constitute a promising point of departure. In addition, several methods exist to compute or approximate classical empowerment in the continuous domain and for high-dimensional problems (e.g., see [4], [8], [38]). Empowerment has also been applied to complex robotic systems (e.g., wheeled robots [13] or autonomous underwater

vehicles [5]). We therefore believe that these approximation methods for empowerment may be a good starting point to expand the GDE formalism towards similar directions. The objective will be to widen the applicability of the GDE framework to the realm of robotics, also bearing in mind the aforementioned empowerment-based HRI applications.

In [31] the use of empowerment to design the behaviour of an artificial agent companion is presented for a video-game scenario. In this study, the non-person character (NPC) companion behaves to maximise the empowerment of the player, resulting in desiderata such as not obstructing the mobility of the player or interfering with its plans when accompanying it and defending the player, eliminating the possible threats that aim at killing it or simply reduce its ability to act. In principle, using GDE the NPC could perform any other task while at the same time increasing the player empowerment. Finally, systems that have to sustain cognitive functions for a prolonged period of time (as those typically found in biology) are current objects of investigation in life-long planning for AI. These systems can face unexpected perturbations that necessitate readiness in order to guarantee a proper reaction. We believe that, through its inherent tendency to foster preparedness with respect to unexpected goals, GDE could be a suitable candidate quantity to select beneficial states for tasks that must be sustained for a extended period of time. We will study that in future work.

REFERENCES

[1] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Keep your options open: An information-based driving principle for sensorimotor systems," *PloS one*, vol. 3, no. 12, p. e4018, 2008.

[2] C. Salge, C. Glackin, and D. Polani, "Empowerment - an introduction," in *Guided Self-Organization: Inception*. Springer, 2014, pp. 67–114.

[3] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[4] T. Jung, D. Polani, and P. Stone, "Empowerment for continuous agent—environment systems," *Adaptive Behavior*, vol. 19, no. 1, pp. 16–39, 2011.

[5] N. C. Volpi, D. De Palma, D. Polani, and G. Indiveri, "Computation of empowerment for an autonomous underwater vehicle," *IFAC-PapersOnLine*, vol. 49, no. 15, pp. 81–87, 2016.

[6] T. Anthony, D. Polani, and C. L. Nehaniv, "On preferred states of agents: how global structure is reflected in local structure," *Artificial Life XI*, 2008.

[7] C. Salge, C. Glackin, and D. Polani, "Changing the environment based on empowerment as intrinsic motivation," *Entropy*, vol. 16, no. 5, pp. 2789–2819, 2014.

[8] S. Mohamed and D. J. Rezende, "Variational information maximisation for intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2015, pp. 2125–2133.

[9] K. Gregor, D. J. Rezende, and D. Wierstra, "Variational intrinsic control," *arXiv preprint arXiv:1611.07507*, 2016.

[10] A. H. Qureshi, B. Boots, and M. C. Yip, "Adversarial imitation via variational inverse reinforcement learning," *arXiv preprint arXiv:1809.06404*, 2018.

[11] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," *arXiv preprint arXiv:1802.06070*, 2018.

[12] F. Leibfried, S. Pascual-Díaz, and J. Grau-Moya, "A unified bellman optimality principle combining reward maximization and empowerment," in *Advances in Neural Information Processing Systems*, 2019, pp. 7867–7878.

[13] A. Leu, D. Ristić-Durrant, S. Slavnić, C. Glackin, C. Salge, D. Polani, A. Badii, A. Khan, and R. Raval, "Corbys cognitive control architecture for robotic follower," in *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. IEEE, 2013, pp. 394–399.

[14] A. Gabriel, R. Akrour, J. Peters, and G. Neumann, "Empowered skills," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 6435–6441.

[15] C. Salge and D. Polani, "Empowerment as replacement for the three laws of robotics," *Frontiers in Robotics and AI*, vol. 4, p. 25, 2017.

[16] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena scientific, 2017, vol. 1, no. 1.

[17] P.-Y. Oudeyer and F. Kaplan, "What is intrinsic motivation? a typology of computational approaches," *Frontiers in neurorobotics*, vol. 1, p. 6, 2009.

[18] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 2, pp. 265–286, 2007.

[19] R. Der, U. Steinmetz, and F. H. Pasemann, "A new principle to back up evolution with learning," *Computational Intelligence for Modelling, Control, and Automation; IOS Press: Amsterdam, Demark*, pp. 43–47, 1999.

[20] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich, "Predictive information and explorative behavior of autonomous robots," *The European Physical Journal B*, vol. 63, no. 3, pp. 329–339, 2008.

[21] R. Der and G. Martius, *The playful machine: theoretical foundation and practical realization of self-organizing robots*. Springer Science & Business Media, 2012, vol. 15.

[22] G. Gordon and E. Ahissar, "Hierarchical curiosity loops and active sensing," *Neural Networks*, vol. 32, pp. 119–129, 2012.

[23] L. Steels, "The autotelic principle," in *Embodied artificial intelligence*. Springer, 2004, pp. 231–242.

[24] F. Kaplan and P.-Y. Oudeyer, "Maximizing learning progress: an internal reward system for development," in *Embodied artificial intelligence*. Springer, 2004, pp. 259–270.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press, 2018.

[26] G. Baldassarre and M. Mirolli, *Intrinsically motivated learning in natural and artificial systems*. Springer, 2013.

[27] N. Chentanez, A. G. Barto, and S. P. Singh, "Intrinsically motivated reinforcement learning," in *Advances in neural information processing systems*, 2005, pp. 1281–1288.

[28] L. Meeden, J. Marshall, and D. Blank, "Self-motivated, task-independent reinforcement learning for robots," *Cybernetics and Systems*, vol. 36, 2004.

[29] J. Schmidhuber, "Formal theory of creativity, fun, and intrinsic motivation (1990–2010)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 230–247, 2010.

[30] H. J. Charlesworth and M. S. Turner, "Intrinsically motivated collective motion," *Proceedings of the National Academy of Sciences*, vol. 116, no. 31, pp. 15 362–15 367, 2019.

[31] C. Guckelsberger, C. Salge, and S. Colton, "Intrinsically motivated general companion npcs via coupled empowerment maximisation," in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2016, pp. 1–8.

[32] D. Barber and F. V. Agakov, "The im algorithm: a variational approach to information maximization," in *Advances in neural information processing systems*, 2003, p. None.

[33] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[34] E. Delage and S. Mannor, "Percentile optimization in uncertain markov decision processes with application to efficient exploration," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 225–232.

[35] K. Regan and C. Boutilier, "Robust policy computation in reward-uncertain mdps using nondominated policies," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[36] H. J. Kappen, "Path integrals and symmetry breaking for optimal control theory," *Journal of statistical mechanics: theory and experiment*, vol. 2005, no. 11, p. P11011, 2005.

[37] R. Fox and N. Tishby, "Bounded planning in passive pomdps," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 2012, pp. 75–82.

[38] C. Salge, C. Glackin, and D. Polani, "Approximation of empowerment in the continuous domain," *Advances in Complex Systems*, vol. 16, no. 02n03, p. 1250079, 2013.