Self-Motivated Composition of Strategic Action Policies

Tom Anthony

August 2017

Submitted to the University of Hertfordshire
in partial fulfilment of the requirements of the degree of
Doctor of Philosophy

Abstract

In the last 50 years computers have made dramatic progress in their capabilities, but at the same time their failings have demonstrated that we, as designers, do not yet understand the nature of intelligence. Chess playing, for example, was long offered up as an example of the unassailability of the human mind to Artificial Intelligence, but now a chess engine on a smartphone can beat a grandmaster. Yet, at the same time, computers struggle to beat amateur players in simpler games, such as Stratego, where sheer processing power cannot substitute for a lack of deeper understanding.

The task of developing that deeper understanding is overwhelming, and has previously been underestimated. There are many threads and all must be investigated. This dissertation explores one of those threads, namely asking the question *"How might an artificial agent decide on a sensible course of action, without being told what to do?"*.

To this end, this research builds upon *empowerment*, a universal utility which provides an entirely general method for allowing an agent to measure the preferability of one state over another. Empowerment requires no explicit goals, and instead favours states that maximise an agent's control over its environment.

Several extensions to the empowerment framework are proposed, which drastically increase the array of scenarios to which it can be applied, and allow it to evaluate actions in addition to states. These extensions are motivated by concepts such as *bounded rationality*, sub-goals, and anticipated future utility.

In addition, the novel concept of *strategic affinity* is proposed as a general method for measuring the strategic similarity between two (or more) potential sequences of actions. It does this in a general fashion, by examining how similar the distribution of future possible states would be in the case of enacting either sequence. This allows an agent to group action sequences, even in an unknown task space, into 'strategies'.

Strategic affinity is combined with the empowerment extensions to form *soft-horizon empowerment*, which is capable of composing action policies in a variety of unknown scenarios. A Pac-Man-inspired prey game and the Gambler's Problem are used to demonstrate this self-motivated action selection, and a Sokoban inspired box-pushing scenario is used to highlight the capability to pick strategically diverse actions.

The culmination of this is that soft-horizon empowerment demonstrates a variety of 'intuitive' behaviours, which are not dissimilar to what we might expect a human to try.

This line of thinking demonstrates compelling results, and it is suggested there are a couple of avenues for immediate further research.

One of the most promising of these would be applying the self-motivated methodology and strategic affinity method to a wider range of scenarios, with a view to developing improved heuristic approximations that generate similar results. A goal of replicating similar results, whilst reducing the computational overhead, could help drive an improved understanding of how we may get closer to replicating a human-like approach.

Acknowledgements

My gratitude extends beyond the limits of my capacity to express it.

— Jernau Morat Gurgeh
The Player of Games (Iain M. Banks)

Daniel

First and foremost, I wish to thank my supervisor, Daniel. My gratitude to Daniel extends further back in time than the start of my PhD.

As an undergraduate I had a plan to write a computer chess engine for my final project and, after several efforts to put me off such a "foolhardy" plan, my project tutor relented and introduced me to Daniel ("if you insist, then I know exactly who might help. . . ").

I remember still, 14 years later, meeting Daniel that the first time – "Chess? That is basically solved! Have you heard of the game Go?". I had arrived in the right place!

I stuck with Chess, and whilst Daniel taught me so very much, the greatest thing he shared with me that year was his infectious passion for science and his enthusiasm that I should pursue it further.

Chrystopher

I recall Chrystopher, my second supervisor, reading drafts in front of me and pointing out, amongst other things, small factual mistakes in the bibliography. It was amazing to me that he had all this information in his head. He helped me understand that any scientist has a responsibility to be precise.

Room E122

Philippe helped me a lot practically, with learning Java and helping me learn how to take my intuition and understanding, and craft it into formulas and equations. Antoine showed me that

science can benefit from a splash of irreverence. Paul taught me that, however I was thinking about something, I should probably try thinking about it differently. Christoph gave me great feedback on various drafts of papers, but I thank him mostly for his dance show.

Sven (HB!) taught me that, when cooking, I should put down the recipe book and pay attention to the ingredients (applies to more than cooking!). He also taught me to have more confidence in my viewpoint. I will forever have great memories of living together (and playing Bubble Bobble for hours on end until we finally beat it).

My Parents

I would never have had the opportunity to pursue this work, or so much else, without the help and support of my family. My parents supported me in many ways and never doubted I could or should do this. I love you, and I'm grateful.

Meine Mädels

I met Sophie in room E122 the day she arrived for a placement from Germany. We gradually became friends, and the more we spoke... the more we spoke. Her relentless persistence eventually wore me down (that is my story and I'm sticking to it). I'll spare the reader from the usual clichés, and instead I'll quote Melvin Udall from *As Good As it Gets*: "You make me want to be a better man".

Since then we have welcomed our two daughters, Mathilda and Philippa, got married, and gone on a number of adventures together. I hope there will be many more to come.

Explaining to Mathilda and Philippa what scientists do (which involves bottle rockets and other messy experiments!) reminds me why being one is important, and fun!

I am so lucky to have you all. I'll try not to forget it.

List of Published Papers

This dissertation is built upon my publications. The following papers form the core of my submission:

Paper 1: Anthony, T., Polani, D., and Nehaniv, C. L. (2008). On preferred states of agents: how global structure is reflected in local structure. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 25–32. MIT Press, Cambridge, MA

Paper 2: Anthony, T., Polani, D., and Nehaniv, C. L. (2011). Impoverished empowerment: 'meaningful' action sequence generation through bandwidth limitation. In Kampis, G., Karsai, I., and Szathmáry, E., editors, *Artificial Life. Darwin Meets von Neumann. ECAL 2009. Lecture Notes in Computer Science*, volume 5778. Springer

Paper 3: Anthony, T., Polani, D., and Nehaniv, C. L. (2014). General self-motivation and strategy identification: Case studies based on Sokoban and Pac-Man. *Computational Intelligence and AI in Games, IEEE Transactions on*, 6(1):1–17

Each paper has a dedicated chapter, which also provides extended discussion and and additional results. The papers are presented here in the same format as they were each published, but with minor edits to correct for typos/clarity, and with margin titles and page numbers consistent with the primary document.

Contents

Chapter 1

Introduction

Introduction

We can only see a short distance ahead, but we can see plenty there that needs to be done.

— Turing (1950)

1.1 Introduction and Motivation

During the Dartmouth Conference in 1956, the attendees searched for the underlying principles that could be used to fully and completely describe the processes that form intelligence. It was soon found that these underlying principles were far more complex, abstract and difficult to identify than had been predicted. AI research has since helped in successfully conquering many smaller problems, and has produced a variety of potent applications. These two branches of research now work towards different goals; with Artificial General Intelligence (Strong AI) working towards the original goals of broad 'human like' intelligence, and Applied AI (Weak AI / Machine Learning) research working on applying AI to smaller, more specific tasks.

However, it seems that the original goal of finding underlying principles has sometimes been pushed aside, and fails to be realised. The story is similar in the parallel field of Artificial Life; the principles fundamental to adaptive behaviour continue to be of interest, but research often adjusts an approach to fine tune it for a specific experiment or application. Such fine tuning can work very well to produce results for a specific niche, but continues to evade the necessary generality that is plausibly the key to simulating biological organisms.

Artificial Life does have an area where this generality exists, namely universal utilities; methods which can be used to evaluate a broad, or even complete, range of scenarios that a would-be adaptive agent may encounter. Biological organisms excel at this, and can evaluate and learn new and complex situations quickly and effectively. Learning more about how universal utilities achieve their results is a pivotal part of understanding what drives adaptive behaviour.

However, we recognise that using computers to dedicate their vast resources to any problem domain is beyond what a biological organism would be capable of doing, even if it sometimes produces similar results. A computer is unlikely to do things *in the same way* a biological organism would do.

In this dissertation I chose to focus on discrete-world problems in which an agent or player has no external goals but does have an explicit model of the dynamics of the world. The assumption is made that most organisms have a reduced model of the physics of their world and their own embodiment, and can understand the likely outcomes (allowing for a level of included stochasticity) of the majority of their actions (with other organisms in the same environment being modelled as part of that same stochasticity).

Scenarios with no predefined goals are of interest, because they demand generality and furthermore tend to highlight the effect of an agent's specific embodiment on how it behaves; a scenario with explicit predefined goals means the pursuit of those goals tends to obfuscate any 'default' behaviours of an agent. It is hypothesised that there is an intrinsic link between an organism's embodiment and the niche in which it has evolved (Pfeifer and Bongard, 2006; Berger, 2003). My work is motivated by the extrapolation of this argument: that the embodiment of an organism can be assumed, from an evolutionary argument, to have been arrived at in order to be well equipped to thrive in its appropriate environment, and that a suitably embodied organism or agent gets both cognitive and metabolic advantages within that environment.

If we were to accept this premise, then it follows that having the appropriate embodiment makes doing the right thing easier (via the reduction in cognitive effort and the increased viability of engaging with relevant affordances). The complement of this argument is that the embodiment of an organism itself can indicate available actions and thus help guide an understanding of what that organism should do.

Put colloquially, an organism, or agent, that has the right anatomy for an environment needs to think less and exert less effort in order to thrive, and its embodiment can give an insight into what behaviours may help it to do so.

Note that the type of scenario discussed above, with an agent with an embodiment relevant to its niche, which lacks specific goals but has an understanding of the dynamics of its environment and the effect its own actions on that environment, is also modelled very well by a variety of game-based scenarios.

1.1.1 Game Playing and Generality

A prominent field of research that provides a look at game-playing scenarios that may be relevant to the outlined motivation is *general game playing* which, along with the wider research into game playing, has made many advances in the last few years. Ever since Garry Kasparov's defeat at the metaphorical hands of Deep Blue (Campbell et al., 2002), there has been increased public interest in the area. More recently, work into Monte-Carlo Tree Search (MCTS) (Enzenberger et al., 2010; Coulom, 2007) has led to very promising results in games such as Go (Silver et al., 2016), which had eluded previous efforts using more traditional tree search algorithms, such as those employed by Deep Blue and similar systems.

A number of variations of MCTS (most notably UCT) have been applied with great success in general game playing scenarios (Mehat and Cazenave, 2008), and the performance of these algorithms is exceptional, especially when considering the limited domain knowledge they require.

However, whilst these algorithms excel performance wise (in terms of results), they are not constrained in the same ways that biological organisms might be. Even in the very recent victory of AlphaGo (Silver et al., 2016) over Go professional Lee Sedol and now Ke Jie, the current world #1, the program made extensive use of MCTS, which the authors describe as "tree search programs that simulate thousands of random games of self-play".

So whilst game scenarios often form a good representation of the type of scenarios we are interested in, the research into general game playing sits on one side of the division between biological plausibility and generality.

On the other side of that division are a variety of proposed universal utility functions and generalised methods (which increasingly include recent developments in ANNs, which AlphaGo also utilised). Of interest to the present research are those based on Shannon-type information, such as *excess entropy / predictive information* (Prokopenko et al., 2006; Ay et al., 2008; Bialek et al., 2001), *homeokinesis* (Der et al., 1999), and the concept of *empowerment* (Klyubin et al., 2005b,a).

In the case of empowerment, previous results indicate that it works flexibly in a range of scenarios with full generality, and that it performs in a way that appeals to an intuitive understanding of how biological systems may approach tasks. However, it was not entirely clear exactly how empowerment was working 'below the surface', and furthermore, the prior research into empowerment also lacks many of the constraints which we would consider necessary for a biologically plausible setting.

Therefore, I wanted to investigate and build upon the established work and research into empowerment to investigate how it works at an underlying level. The hypothesis, also related to the observations in the previous section, was that it is the embodiment of an agent or organism that enables it to utilise the underlying structure within an environment to assess the desirability of certain states. Furthermore, by helping develop a better understanding of some of the mechanisms that allow such a parsimonious method of 'general utility' estimation,

I hoped that we may gain a better understanding of how it operates within the constraints that it does.

It seemed clear that nature has better heuristics for achieving similar behaviours to those generated by empowerment, and that it does so whilst operating under its own set of constraints. I hoped that further constraints on empowerment may move us towards a better understanding of those heuristics, and help develop novel approaches for adaptive systems.

1.1.2 Considerations

Following from the above, my work is based on an assumption that in order to discover biologically plausible universal utilities, it is more important to operate within similar constraints to those of biological organisms, rather than to focus on performance (both in terms of processing and of results).

If the primary measure of early success in the field was performance, then there is an incentive for any technique to both specialise in a specific niche, and to leverage the sheer computational power available from computers.

In summary, the considerations within my thesis are:

- the agent should operate within constraints inspired by biological organisms;
- the utility function must arise from a generic model, and not be targeted towards any specific problem;
- performance of the agent, both in terms of processing speed and of results, is less important than operating within the constraints above.

Of course, it is impossible to start with all the constraints of a biological system for several reasons: they are often unknown and variable, and to do so would confine us to small domains that would restrict our ability to evaluate any results. The aim here has to been to find a balance

between some approximate constraints which would be broadly universal, and working with sufficiently non-trivial scenarios to evaluate performance.

Furthermore, with a discussion of universal utilities, especially in unknown scenarios, including games without explicit goals, we must quickly discuss the role of exploration versus exploitation. Throughout my papers, whilst making some speculation about exploitation, the focus is primarily on the exploration phase, where the environment is new and there is not yet a clear goal.

It is clear that biological organisms learn and adapt to specific scenarios as a key part of surviving and prospering, which is more akin to exploitation; the work presented serves as a good basis for learning and adaptation but focuses on general methods rather than scenario specific methods, more akin to exploration.

1.2 Contributions of the Dissertation

The work presented in this dissertation makes the following contributions:

Applying information bottleneck technique to perception action loop: Due to its nature, calculating n-step empowerment previously required examining all possible action trajectories for an agent through the state space. This made it computationally inviable to calculate empowerment for sequences longer than a few steps in most scenarios. However, any method to overcome this needed to fit within the same information theoretic framework and furthermore should be capable of retaining full empowerment (i.e. not result in a reduction in empowerment). Presented initially in Paper 2, the method uses the information bottleneck method, which is capable of entirely removing redundancy from an action repertoire; this reduces the space required to retain a 'fully empowered' action repertoire, and combined with the 'iterative extension of the empowerment horizon', provides a more tractable method of calculating empowerment.

Iterative extension of the empowerment horizon: Building upon the improved empowerment methodology that uses the information bottleneck outlined above to reduce the number of retained action sequences, a method was developed of iteratively extending the retained action sequences such that the viable horizon for empowerment is extended by orders of magnitude.

Applying empowerment to actions: Previously empowerment had provided a general utility for evaluating states, and comparing them. Despite empowerment being about what an agent or organism can *do*, there existed no method for evaluating actions, and which actions were most responsible for providing empowerment. The 'compression' provided by the application of the information bottleneck method to empowerment also provided a method for evaluating which actions contributed towards empowerment. This method helped drive self-motivated action selection, by selecting a subset of actions. The initial work provided a naive approach to this but this was improved with the introduction of *soft-horizon empowerment*.

Soft-Horizon Empowerment: In working to apply empowerment to actions it was recognised that, after identifying redundancy amongst action sequences, the most important facet of understanding which actions were more empowering was actually considering the empowerment of the state or states the agent would reach should you enact that action (i.e. did an action sequence lead somewhere good). To address this, the concept of a second horizon within the same model was introduced, allowing the comparison of actions with one another with regards to the future empowerment they might provide. This improved the self-motivated action selection driven by constraining the retainable action sequences towards selecting actions that led to ongoing future empowerment. In addition, Soft-Horizon Empowerment also utilised the *strategic affinity* to diversify the action sequences selected, as discussed below.

Strategic Affinity: Having extended empowerment with a concept of compression to drive action selection, iteratively extending the horizon to improve performance and generality, and a second horizon to guide action selection toward more empowering actions, there was a need to now drive towards selecting a richer and more diverse set of actions. I proposed the novel concept of *strategic affinity*, which measures the similarity or 'overlap' of reachable future states between various action sequences. This allows those action sequences to be clustered according to how similar the future looks should they be enacted. In Paper 3 I present this concept entirely within the framework of empowerment and information theory, but the concept may be applicable to other methods.

1.3 Outline of the Dissertation

The body of this dissertation is built upon a set of published papers which have been peer reviewed. Each paper makes a contribution to the thesis, building upon one another and leading the research from an analysis of environmental structures and how they indicate preferred states (Paper 1), through to how to extend this from states to actions, and thus produce behaviours (Paper 2), and finally applying the novel concept of 'strategic affinity' to refine these behaviours to produce strategically diverse repertoires of actions (Paper 3). Initially an introductory chapter introduces empowerment. A final chapter discusses the contributions and future work.

A short summary of each chapter:

Chapter 2 introduces empowerment, its links to other theories and some previous results. The chapter identifies some problems with empowerment, discusses some hypothetical new applications for it and describes how it seems to leverage underlying structures in an environment.

Chapter 3 presents Paper 1. To begin I present the motivation for the paper, focusing on the hypothesis that an agent or organism's embodiment can give indications to that agent or organism's preferred states in the world. Following the paper the main directions for the work to follow are outlined: better differentiation of a state's utility without requiring prior knowledge of the best 'horizon', and determining how to evaluate empowerment of actions.

Chapter 4 presents Paper 2. The paper is preceded by a discussion around the links between compression and intelligence, and an extension of the hypothesis around embodiment indicating preferred states applying to actions. These topics are linked together with the previous paper as an introduction to the next. After the paper, the discussion looks at addressing the weaknesses - namely finding a method of self-motivated identification of strategically diverse actions.

Chapter 5 presents Paper 3. In addition to the results in the paper, additional results are also presented from the Gambler's Problem. The discussion section outlines how a second horizon can help drive more salient action selection. Further, a comparison between soft-horizon empowerment and greedy mobility climbing is analysed, and the surprising kiting behaviour is discussed. Finally, the language around the methods presented in the paper, including strategic affinity, is standardised.

Chapter 6 summarises the main contributions from the papers, and adds some additional commentary. Finally, I present some ideas about the future direction of work, both in regard to the specific results presented throughout the dissertation, and also the wider goals that motivated the work.

Chapter 2

Background and Empowerment

Background and Empowerment

I shall act always so as to increase the total number of choices.

— Von Foerster (2003)

2.1 Motivation

Most organisms cannot detect every facet of their environment, detecting only certain wavelengths of the electromagnetic spectrum, or certain frequencies of sound, but with better vision, hearing, or other senses (for 'free'), they could likely perform better in their niche. There is obviously a metabolic cost to sensing, and it seems intuitive that the process of evolution drives successive generations of a species towards an advantageous trade off point between minimum metabolic cost and maximum perception utility of relevant aspects of their environment.

Therefore it is interesting to determine which facets of the environment hold the most utility, and understand why this might be. Gibson noted that organisms do not view their environments as a geometric space but that "the theory of affordances implies that to see things is to see how to get about among them and what to do or not do with them" (Gibson, 1979), and so it is not unreasonable to hypothesise that this in turn implies that evolutionary pressures would drive successive generations of an organism towards having sensory apparatuses that detect as many observable affordances as is viable.

However, this is still not a complete picture, because not all affordances are relevant to an organism's survival. If we consider again the metabolic cost of sensors, combined with Gibson's observations, then we could understand that the process of evolution has provided an organism with those sensors that represent a trade-off between its perception of affordances relevant to its survival and the metabolic costs of those sensors.

Put differently, organisms have sensory apparatus that allow them to see what they can *do* in their environment, and those actions likely have some utility for the organism.

We can apply this same reasoning to an organism's motor apparatus, and reach the conclusion that those too have evolved to allow the organism to perform the actions (or allow the affordances) most relevant to its survival.

So let us work on the basis that an organism or agent's sensorimotor apparatus has been adapted for perceiving the environmental cues, and performing the actions, that are most relevant to its success. Now we can postulate that, all else being equal, we might expect it to perform better in situations that provide more opportunities to make use of that apparatus.

Likewise, if we remove an organism from the niche for which it evolved, and transferred it to a dissimilar environment, we would expect the organism to perform poorly. This is the principle that sets the foundation for the hypothesis of empowerment.

Introduced in Klyubin et al. (2005a, 2004a, 2008):

> *Empowerment is the* perceived amount of influence or control *the agent has over the world. For example, if the agent can make one hundred different actions but the result,* as perceived by the agent*, is always the same, the agent has no control over the world whatsoever. If, on the other hand, the agent can reliably force the world into two states distinguishable by the agent, it has two options and thus two futures to choose from. Empowerment can be seen as the agent's* potential *to change the world, that is, how much the agent could do in principle.*

Put succinctly, empowerment is a measure of an embodied agent's perceived utility in an undefined task space.

Empowerment uses Shannon's information theory (Shannon, 1948) combined with the perception-action loop formalism to model the interactions between an agent and its environment as a communications channel.

Empowerment does not (at least in its original form) suggest or imply which actions an agent should necessarily perform, but rather provides a general measure of utility that could be used to compare one state to another.

There are obviously a variety of possible ways in which the concept of 'prefer states with more options' could be modelled and approached, but in choosing an information theoretic basis, Klyubin et al. ensured that the surrounding model stayed free of any semantic or symbolic meaning (Klyubin et al., 2005a). This is important for several reasons:

- Shannon style information can model any environment in an entirely general fashion, which extends to an agent's interactions with that environment, which minimises the chance of inadvertently introducing an assumption or bias into the model;
- in turn this highlights that empowerment does not rely on any facets of the underlying framework on which it is built;
- the generality and the 'toolbox' provided by the information theoretic framework ensure the model is accessible to analysis.

Empowerment is modelled upon the perception-action loop, an information theoretic model first introduced in Klyubin et al. (2004b) and inspired by Touchette and Lloyd (2004), which applied Shannon's information theoretic model of communications to control theory.

2.2 Information Theory

Information theory was introduced by Shannon (1948), and is well suited for modelling and processing an embodied agent's perception of its environment and internal memory. In searching for a formalism that is both biologically plausible and universal, and which can be well understood, it is clear that Shannon's model is a strong candidate.

Information has the advantage of being elementary in nature; Shannon's model did away with semantics and presented a 'pure' unit, which exists everywhere, and can record anything that an organism, or agent, can perceive. Building adaptive behaviour from the ground up using information theory will ensure formalisms which are general, and not specific to one model or another.

To begin we introduce *entropy*, which is a measure of uncertainty:

$$H(A) = -\sum_{a \in A} p(a) \log p(a). \tag{2.1}$$

where A is a discrete random variable with values $a \in A$ and $p(a)$ is the probability that A assumes a. The logarithm can be taken to any chosen base; in this dissertation and the included papers I consistently use 2, and accordingly the units of measurement are then called *bits*. If S is another random variable jointly distributed with A, the *conditional entropy* measures the remaining uncertainty about the value of S, if we know the value of A:

$$H(S|A) = -\sum_{a \in A} p(a) \sum_{s \in S} p(s|a) \log p(s|a). \tag{2.2}$$

Its relationship to entropy is shown in Fig. 2.1, which allows us to measure the *mutual information* between two random variables:

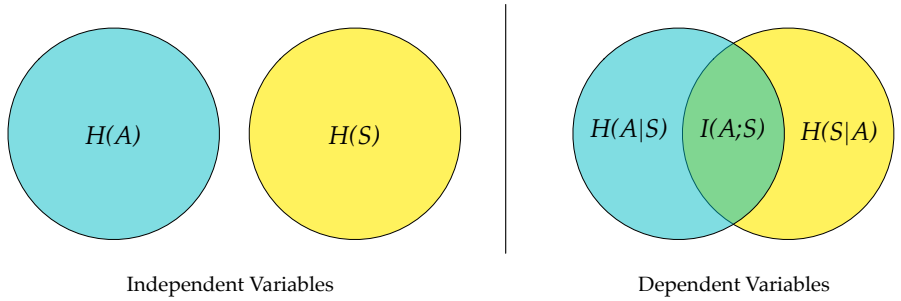


Figure 2.1: *Visualization of the primary information theory quantities and their relationships. $H(A)$ is the entropy of the random variable A, $H(A|S)$ is the conditional entropy of the same variable conditioned on knowing the value of S. Finally $I(A;S)$ is the mutual information between these two variables, which measures the amount of information that the two variables reveal about one another.*

$$\begin{aligned} I(A;S) &= H(S) - H(S|A) \\ &= \sum_{a\in A}\sum_{s\in S} p(a,s)\log\left(\frac{p(a,s)}{p(a)\,p(s)}\right) \end{aligned} \tag{2.3}$$

Mutual information can be thought of as the reduction in uncertainty about the variable A or S, given that we know the value of the other. The mutual information is symmetric, so we could also use $I(A;S) = H(A) - H(A|S)$ (Cover and Thomas, 1991). In Paper 3 we will later also examine the mutual information between a single value of a random variable with another random variable:

$$I(a;S) = p(a)\sum_{s\in S} p(s|a)\log\left(\frac{p(a,s)}{p(a)\,p(s)}\right). \tag{2.4}$$

Channel Capacity

Finally, we introduce channel capacity (Shannon, 1948; Cover and Thomas, 1991) for a discrete memoryless channel, defined as the maximum mutual information over the channel for all possible input distributions:

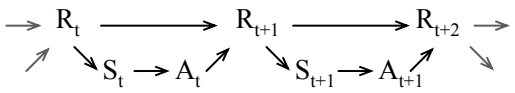


Figure 2.2: *Bayesian network representation of the perception-action loop.*

$$C(p(s|a)) = \max_{p(a)} I(A;S). \tag{2.5}$$

The random variable A represents the distribution of messages being sent over the channel, and S the distribution of received signals. Clearly, the higher the mutual information between the two variables, the higher the capacity of the channel. The channel capacity is measured as the maximum mutual information taken over all possible input distributions, $p(a)$, and depends only on $p(s|a)$, which is fixed. One algorithm that can be used to find this maximum is the iterative Blahut-Arimoto algorithm (Blahut, 1972).

2.3 Empowerment

2.3.1 Perception Action Loop

Empowerment is a measure of the efficiency of a *perception-action loop*. The model of the perception-action loop allows the flow and processing of information to be easily understood and modelled, and has proven to be an effective foundation for modelling adaptive behaviour.

In Fig. 2.2 we can see the perception-action loop represented by a causal Bayesian network, where the random variable R_t represents the state of the environment, S_t the state of the sensors, and A_t the actuation selected by the agent at time t. Here we assume that R_{t+1} depends only on the state of the environment at time t, and the action just carried out by the agent.

This can be looked upon as the agent 'injecting' information into the environment through its actuators, and receiving information from the environment through its sensors. So it

can be thought of as a communication channel, and by modelling it as one we can employ information-theoretic methods, which are the basis for empowerment.

2.3.2 Calculating Empowerment

It is likely that an agent performing an action in any moment is likely to change the state of the world, and in this model, that is represented in Fig. 2.2 by a flow of information from A_t into R_{t+1}. The state of the world has now changed, and this change can be perceived by an agent's sensors in the subsequent time step, represented by S_{t+1}, which completes the flow of information throughout the loop.

It may be that the flow of information between any two of these variables may or not be complete at any point; it might be that an agent's action does not affect the environment, or simply that the change is not wholly perceivable by the agent's sensors (e.g switching a switch that turns on a light in another room; the agent can perceive the new state of the switch but not detect the changed state of the light).

Representing the relationship between an embodied agent and its environment in this way allows an array of information-theoretic methods to be used in its analysis. Empowerment, being based upon channel capacity, as a measure of the maximal flow of information over a channel, is usually measured in *bits*. It can be defined as:

$$\mathfrak{E} = C(p(s|a)) = \max_{p(a)} I(A;S). \tag{2.6}$$

In this way empowerment can measure an agent's ability to 'transmit' information to itself from its actuators, A_t, to its sensors,S_{t+1}, via the environment, and can identify the distribution of input signals that would maximise the potential flow of information over this channel. This corresponds to an agent's ability to cause observable change in its environment, and can suggest to an agent the action policy that would maximise this potential to affect change.

Obviously, an agent's ability to cause change in its environment is based on the agent's current state within that environment, and can be measured in various states in order to discover the most empowered. However, in a single time step it would normally be quite difficult to distinguish a difference in utility between two quite different states, and so the concept of *n-step empowerment* needs to be introduced.

2.3.3 *n*-step Empowerment

With *n*-step empowerment, rather than measure the potential flow of information from an agent's actuators back to its sensors via the environment in a single time step, we instead measure the flow of information, aggregated over a sequence of actions, to the subsequent time step. The length of this sequence, n, defines the *horizon* that we measure empowerment for.

Formally, we first construct a compound random variable of the next n actuations $(A_t, A_{t+1}, A_{t+2}, \dots, A_{t+n}) = A_t^n$. We now maximise the mutual information between this variable and the state at time $t+n$, represented by S_{t+n}. n-step empowerment is the channel capacity between these:

$$\mathfrak{E}^n = C(p(s_{t+n}|a_t^n)) = \max_{p(a_t^n)} I(A_t^n; S_{t+n}). \tag{2.7}$$

Fig. 2.3 demonstrates using n-step empowerment to measure the utility of states in a simple 2-d gridworld maze example; in this example the agent's actions are simply to be able to move in the 4 cardinal directions.

Using channel capacity, and based on the idea that 'keeping your options open' is good, empowerment can be thought of allowing agents to understand which situations maximise its potential to utilise its sensors and actuators. In a biological context, this could be justified by the fact that empowerment has equipped an agent or organism with the sensorimotor appa-

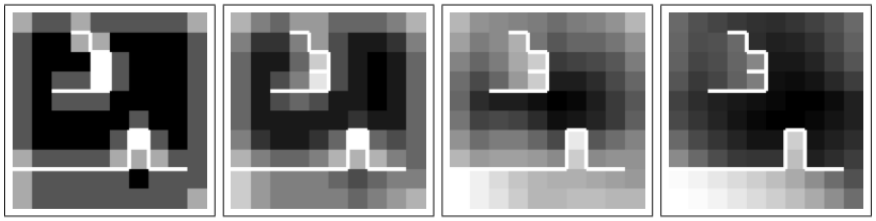


Figure 2.3: *1, 2, 5 & 10-step empowerment for a 10x10 maze (Klyubin et al., 2005a). Darker areas represent higher empowerment.*

ratus that allows it to succeed in its niche, and so maximising the use of those sensors is a sensible, albeit naive, policy from an evolutionary point of view.

We note that *n*-step empowerment requires a 'god view' of the world, in that it requires an understanding of the distribution if future outcomes; we will discuss this problem more in Section 2.7.2.

Throughout this dissertation I will sometimes use 'action' in place of the more complete 'action sequence', where it is clear, for the sake of readability.

2.4 Underlying Environmental Structure

Empowerment provides a powerful and flexible method for guiding agents to preferential points in an environment, and achieves this at a very low level. A variety of results corroborate its efficacy, and high empowerment states are anecdotally consistent with those somebody would intuitively choose. For example, Fig. 2.3 shows a maze, with each cell shaded to show n-step empowerment, with darker areas having higher empowerment. It is evident that areas which allow access to most of the maze in the shortest number of steps should have higher empowerment, and the diagram clearly demonstrates that to be true.

This tendency for empowerment to align with 'native' measures in any particular task spaces was not unexpected, but it was not fully understood how empowerment was able to correspond so well in unknown environments.

Figure 2.4: *The same maze as in Fig. 2.3, now coloured to show average shortest distance to all other cells in the maze (Klyubin et al., 2005a).*

It is easy to see that an environment or task space that is entirely random would be essentially impossible to operate effectively within; if there is no underlying structure then nothing can be manipulated in a predictable manner. The same logic applies to the specific processes of learning, adaptation and evolution. For empowerment to achieve these empirical results, it must be 'in touch' with this structure.

Empowerment operates solely by implicit analysis of the potential information flow through the perception-action loop - information that we know represents an agent's sensoric perception of the environment. This information encodes the state of the environment as it changes over time, and will capture whatever aspects of the environment's structure that are accessible to the agent.

A motivating hypothesis of this thesis was that the underlying structure of an environment is essential to empowerment's success, and in looking to understand empowerment better and how it might perform when constrained, it was important to understand how empowerment interacted with this structure.

Chapter 3 looks at the structures that are underlying within certain environments, their relationship to empowerment, and links to related types of structure.

2.5 Links to Other Theories

There are a number of other theories which are interrelated with the theory of empowerment, some of which are higher level principles related to biology and artificial life, and others which are more specific but share similar motivations.

Here I present a few theories which are particularly relevant to this thesis. I have broadly divided them into three groups: gameplaying and decision making, biologically grounded methods, and information theoretic methods.

2.5.1 Gameplaying and Decision Making

Decision Theory / Game Theory

Normative decision theory deals with identifying the optimal decision for a rational agent in a given scenario, where that scenario may involve a degree of uncertainty. Commonly the focus is on the *expected value* (Hamming, 1991) of a certain decision or set of decisions. There are obvious parallels here with empowerment which is also a utility measure and which can also operate in stochastic scenarios.

However, decision theory deals with specific scenarios on a case by case basis, whereas empowerment attempts to be fully general.

In game theory, the aims and methods are similar to decision theory, but with the introduction of other agents into a scenario, which are frequently antagonists of the decision maker. Empowerment fully handles this extension without modification, as other agents can be modelled within the channel with no other modifications to the algorithm required.

Bounded Rationality

When considering biological constraints, decision theory introduces the concept of *bounded rationality*, whereby an agent must make decisions under cognitive constraints. Bounded rationality is discussed further below, in Section 2.5.2.

St. Petersburg Paradox

The St. Petersburg paradox (Eves, 1990; Weiss, 1987) is a well known paradox in decision theory & economics, where a player is offered the chance to take part in a coin-flipping game. The game is simple: a coin is tossed and should it come up tails then the player wins \$2, but if it comes up heads then the prize is doubled to \$4 and the game continues. The game goes on like this doubling the prize for each heads until the first tails result is encountered.

The question is how much should a player choose to pay to take part in this game? A straightforward mathematical analysis would suggest that the expected value is infinite:

$$\begin{aligned} E &= \frac{1}{2}\cdot 2 + \frac{1}{4}\cdot 4 + \frac{1}{8}\cdot 8 + \frac{1}{16}\cdot 16 + \cdots \\ &= 1+1+1+1+\cdots \\ &= \infty \end{aligned} \tag{2.8}$$

However, whilst the expected value is infinite, any player who bet all they had to play would end up bankrupt. The calculated expectation is contingent on an infinite sequence of heads, which is where this discrepancy appears.

Various different proposals have been made as to how one can resolve the paradox, one of which is the *expected utility hypothesis* introduced by Bernoulli (1738). Bernoulli's argument is that there is a diminishing marginal utility to money, and that each increase in wealth provides fewer new opportunities. Therefore, he argues, the player should adjust the amount they are

willing to pay based on their current wealth, and he proposed some formulas to calculate what that amount should be.

We can identify that this principle is well aligned with empowerment, and that a theoretical version of empowerment that could calculate over an infinite horizon, might identify that at some point additional money provides no more empowerment, thus providing a hard cut-off in the game beyond which additional winnings are of no benefit.

Such a theoretical empowerment does not exist, but a similar gambling problem was introduced in Dubins and Savage (1976), in which a player must calculate the optimal bet to win a similar coin-tossing game. In section 5.3.1 of Chapter 5 empowerment is used to calculate a betting strategy for that scenario.

General Game playing

Since 2005 Stanford have sponsored a competition into *General Game Playing* (Genesereth and Love, 2005), which they define as:

> *A general game playing system is one that can accept a formal description of a game and play the game effectively without human intervention. Unlike specialised game players, such as Deep Blue, general game players do not rely on algorithms designed in advance for specific games;*

The Stanford competition comes with a specific set of rules and constraints, but one may consider the wider research into general game playing (with a definition essentially the same as the quote above).

Much of the research into the GGP tournament has focused on Monte Carlo tree search (Mehat and Cazenave, 2008; Bjornsson and Finnsson, 2009; Świechowski and Mańdziuk, 2015).

General game playing is of interest to the current dissertation, as it has obvious links to the idea of universal utility functions. Games offer a variety of well defined scenarios which are usually easy to evaluate performance in, and therefore make for an attractive framework within which we can research universal utility functions, such as empowerment.

2.5.2 Biologically Grounded Methods

Perceptual Learning

In general, the existing research around empowerment assumes that the sensory apparatus of any agent has been selected over time by evolutionary pressures. We can imagine that in such a long-term scenario that the principles of empowerment could be a useful method of partially guiding such evolution.

Essentially, this would involve developing sensors by way of maximising the empowerment achievable with them which, when combined with survival pressures, could lead to interesting behaviours. Indeed this concept is touched upon in Klyubin et al. (2008), where sensors are evolved in a gridworld scenario in order to maximise empowerment.

The field of perceptual learning is related, and deals with agents actively improving their perceptual abilities by learning to better differentiate the inputs.

In perceptual learning, the physical sensor does not usually change, but instead an agent learns to better understand the signal from that sensor. Essentially the cognitive layer of an agent learns to better differentiate and interpret that signal.

The current research does not touch on this aspect, but I believe there are definitely interesting avenues of research that could take some of the ideas presented here and apply them as a driving mechanism in making those cognitive improvements.

Affordances

Gibson introduced the concept of affordances in Gibson (1977) and expanded on it in Gibson (1979). There is a clear link between the concept of affordances and that of empowerment, in that they both encapsulate the relationship between an agent and its environment, and the ways in which they can interact.

An affordance is a property not of an agent or the environment, but of the pair of both agent and environment taken together, and recognises that affordances within an environment can differ for different agents. This is similar to empowerment which relies on both the environment, and the embodiment and available actuators of an agent.

Gibson's definition of affordances was considered incomplete and contradictory, and there were several attempts to standardise the definition of affordance including Reed (1996). Reed pointed out "at the general or abstract level, what counts as a given affordance does not have to be the same from one species to another" and gave as an example, "the surface of a pond affords walking-on for small insects, such as water boatmen, but not for any mortal human".

Affordances should normally result in an increase in empowerment; for example, the water boatman can move to many locations by walking on the surface of a lake whereas the mouse is unable to do so. The mouse receives no affordance from the water, which is mirrored by a limited increase in empowerment (albeit a slight one - from the ability to sink!) compared to the water boatman.

However, there are also some important differences between the two theories. Empowerment requires that an agent is capable of observing the change in state that may come about from a specific action in order for it to result in an increase in empowerment, whereas an affordance exists independently of an agent's ability to perceive its presence, or any change in state that may result from it being used. An affordance, which is available to an agent, but for which the outcome of being utilised cannot be observed (e.g. a button that opens a door in another room which the agent is unaware of) would not provide any increase in empowerment.

An interesting observation from Gibson (1977), which relates closely to empowerment, is “I suggest that a niche is a set of affordances.”. This idea that a niche is defined by the relationship between an agent, an environment, and their ability to interact with one another, is related to the hypothesis that organisms have evolved to have the optimal set of sensors and actuators for their niche, which allow them to effect the set of changes in their environment that will potentially maximise their utility.

Furthermore, another observation from Gibson (1977) is that “The niche implies a kind of animal, and the animal implies a kind of niche". The idea that “the animal implies a kind of niche” is parallel to the hypothesis that an agent’s embodiment might indicate preferred states or actions, which is present in Chapters 3 and 4.

Bounded Rationality

In determining what constraints might be reasonable to assume for a biological organism, specifically those regarding memory and computation, there are obvious parallels to the concept of *bounded rationality* (Simon, 1957; Williamson, 1981), which deals with decision making when working with limited information, cognitive capacity, and time. Bounded rationality is used as a model of human decision-making in economics (Tisdell, 1996).

There are various aspects of biological cognition which can be successfully described and understood by assuming informational processing costs being imposed on organisms (Attneave, 1954; Barlow, 1959, 2001; Atick, 1992; Prokopenko et al., 2006; Bialek et al., 2001; Ay et al., 2008; Ortega and Braun, 2011). Furthermore, there are reasons to suspect that the ability of biological cognition to structure its decision-making process is driven by the necessity to economise its information processing (Laughlin et al., 1998).

Applying suitable bounded rationality assumptions to universal utilities could introduce such a necessity to economise information processing, and will not only make them more feasible as models of biological systems, but drive them to generate structured behaviours.

Within the framework of empowerment these types of constraints can be entirely represented in terms of an information-theoretic framework, which is demonstrated in Papers 2 & 3.

2.5.3 Information Theoretic Methods

Infotaxis

Infotaxis is an information theoretic method of searching, that mimics how organisms may conduct an olfactory search for a source that is dispersing sparsely in some medium (Vergassola et al., 2007). Whereas chemotactic bacteria, conducting a similar task, are able to make use of local concentration gradients to guide them towards the source of a nutrient, such a local gradient is not available in many scenarios. In a flowing medium, such as air or water, regions of high concentration are broken into random and disconnected patches.

In such examples, a searcher will only get intermittent and partial cues as to the presence and possible location of the source, and from these local cues must make global conclusions. There is a similarity here with some of my work presented in Paper 1, where an empowerment approach is used to try to infer global structure from local information.

An instance from nature of such an example is the approach of a moth that searches for a mate via an olfactory search for pheromones, which requires a search approach able to operate based on sparse 'hits' in a turbulent medium. The search patterns of moths are well understood, and they "are known to proceed upwind by way of counterturning patterns of extended ('casting') or limited ('zigzagging') crosswind width" (Vergassola et al., 2007).

Infotaxis provides an information theoretic searching algorithm that produces similar behaviours to the moth searches, but is based on information theory.

Chemotactic searches are based on gradual acquisition of information on the source location, which it does via concentration, which gets stronger as the gradient is followed. Infotaxis works similarly, working to reduce the informational uncertainty by locally maximising the expected rate of information gain.

Whereas the goals of infotaxis are clearly very different to those of empowerment, it highlights that information theoretic based methods can produce biologically plausible behaviours. Whilst it is unlikely that moths are computing entropy as part of their search methodology, the information theoretic approach can provide an approximation of the same behaviours, working towards helping develop a better understanding. Through further study of such an approach, it would perhaps be possible to produce heuristics that produce similar results with far less computation required.

With empowerment we see a similar situation, the results seem to align well with intuitive approaches and possibly biologically plausible methods, but it is unlikely that such a computationally heavy mechanism is used in nature. I believe that future work on empowerment should also try to produce heuristics that provide similar behaviours. However, the parallels between infotaxis and empowerment lend weight to the ability of an information focused approach producing biologically plausible behaviours.

Predictive Information

Predictive Information, introduced in Bialek et al. (2001), is the mutual information between the future and the past. Put informally, it measures how much can the past help you predict about the future, and was presented as a measure of complexity for a time series.

It was first applied to robotics in Ay et al. (2008), where it was proposed as a method of self-organisation for an agent's behaviour. Whilst the agent's interaction with the environment impacts the complexity of the time series of the agent's sensor values, it can then be used as a utility function that can help drive the behaviour of the agent.

In Ay et al. (2008), the authors present a two-dimensional 'maze'-type example with a wheeled-robot and a set of obstacles. The robot has two-wheels, which are actuators, from which it can detect rotational velocity and, importantly, it has no proximity sensor. It is capable of detecting the objects only through collisions which change the the behaviour of the wheels.

The results in Ay et al. (2008) demonstrate an interesting effect that, with correct parameter values, the robots develop an exploratory behaviour whereby they explore the space in a self-motivted fashion.

Further work in Der et al. (2008), demonstrated a co-ordinated behaviour amongst a chain of coupled robots, which enables an exploratory behaviour despite no central control.

The work has obvious parallels with Empowerment; it is based on an information theoretic framework, and optimises over the sensorimotor loop of an agent as a form of channel, albeit in a different fashion.

Furthermore, the mutual information between the past and the present has links to the concept of maximising ones options, as per empowerment, but at the same time also runs counter to it. Empowerment maximises the potential different states that one can reach, which includes states that may be unexpected, or represent a dramatic change in state, which may represent a reduction in predictive information.

Homeokinesis

Homeokinesis has the stated aim of helping develop a 'curious and self-exploratory robot', that uses two neural networks to control the robot and drive learning (Der et al., 1999; Der and Martius, 2011). The first network is used directly as the controller, and the second is used to build a model of the world such that it can predict sensor and motor states.

In many traditional setups the approach would involve optimising to minimise the predictive error of this second network, such that the world model most accurately predicts the following state. However, that often leads to a scenario where the agent reverts to doing nothing, such that the world is maximally predictable.

In homeokinesis the aim is drive exploration and curiosity, so the approach is slightly different, and is designed to drive action. So instead, the predictive error is replaced with the

time-loop error, which is essentially the difference between the sensor values at the previous time step and the sensor values at the previous time step as predicted from the current sensor values by the world model. In Der and Martius (2011), the authors describe it as the 'time inverted counterpart' for predictive error; the time-loop error can also be called the *reconstruction error*.

This approach drives an agent towards activity, as the authors explain:

> *When using a retrospective model the system is found to develop this drive for activity by itself due to spontaneous symmetry breaking mechanisms known from the physics of self-organizing systems. The reason is that the model now generates a dynamics backward in time leading to fluctuation amplification which is a necessary prerequisit for self-organization.*

Whereas homeostasis regulates towards stability, the result of a homeokinesis approach is that agents produce stable kinetic behaviours. Examples include wall following behaviours, ball balancing, and navigation through corridors.

Homeokinesis drives towards self-motivated action, which aligns it with one of the motivations of this dissertation. However, there are some key differences between the results of homeokinesis and those presented here. One is that homeokinesis, after a period of adaptation to a new set of sensory inputs (e.g. being moved), produces a single stable action which changes only if the sensory values change. For example, in a circular world with a wall, the agent will develop a wall following behaviour which is entirely stable until the agent is moved farther from, or nearer to, that wall, at which point the robot will adapt to a new stable action. Empowerment, on the other hand, is about managing a wealth of potential different actions.

Figure 2.5: *An example showing scaled empowerment values for $n = 5$ in a simple maze example. A darker shade indicates higher empowerment, which correlates well with the average shortest path from each cell to all others. Empowerment values in this example range from $\mathfrak{E}$=2.58 to $\mathfrak{E}$=4.64 bits.*

2.6 Previous Results

Here I quickly recap some of the previous results of empowerment which are particularly relevant to this dissertation.

Maze

This scenario consists of a typical maze scenario in a 2-D gridworld, through which the agent can move horizontally or vertically with its movement constrained by walls. With one north, east, south, west movement per time step, empowerment was measured for each cell in the gridworld for *empowerment horizons* given by n (Klyubin et al., 2005a). The results showed that by only considering a single time step ($n = 1$), the empowerment values are still quite crude, but as the horizon is extended it can be seen that empowerment is increasingly able to differentiate the preferable (in terms of mobility) states from the less preferable ones.

The important observation is that the empowerment values for a cell correlate well with the length of the average shortest path from this cell to all others; more generally, it correlates strongly with the graph-theoretic measure of closeness centrality (Anthony et al., 2008). This is of interest, since the centrality measure has been designed specifically for graphs, while empowerment can be used in a more general context of (possibly stochastic) actions of generic agents in an environment.

Box Pushing

We consider an agent in a gridworld with the same action set, but with the walls removed. Instead, a box is introduced which the agent is able to manipulate by pushing: when the agent moves from a square adjacent to the box into the cell occupied by the box, it causes the box to move in the same direction by one cell. The agent is able to detect the location of the box and thus detect the state change when it moves, but all states are equal and thus there is no explicit reward associated with moving the box.

Here, empowerment indicated a preference for the agent to be closer to the box. Far from the box, the actions of the agent caused only the latter to move, but close to the box, the agent's actions could additionally manipulate the box's position, leading to a significant increase in empowerment. The bottom line of this experiment is that empowerment identifies manipulable objects, or, more generally, manipulable degrees of freedom (within the state space).

Pole Balancing

As another class of scenarios, empowerment was considered in the pole-balancing task often encountered in control theory (Klyubin et al., 2008). The scenario describes a pole balancing upright upon a cart, which moves on an even surface. The agent in this scenario has two actions: applying an always equal force to either move the cart forwards or backwards. Traditionally, the aim is to move the cart forwards and backwards as necessary to keep the pole balanced, starting from some initial angle. Should the pole move too far in one direction, then it will exceed a threshold making it impossible to recover and will fall flat.

In the empowerment-based approach, no external reward was used; instead, empowerment, which derives directly from the intrinsic system dynamics, was utilised to guide the system's behaviour. Empowerment identified the pole being perfectly upright as amongst the

most empowered states, and at progressively steeper angles, empowerment drops, until it falls to 0 when the pole becomes utterly uncontrollable.

Using an action selector that greedily maximises predicted empowerment in the following time step, leads to pole-balancing. This observation generalises to other, more complex balancing scenarios (Jung et al., 2011).

2.7 Problems with Empowerment

Empowerment provides an alternative, biologically motivated, method that was shown to do well in various problem domains using a very generic approach. However, there were some notable limitations with the framework that had previously prevented it being suitable for all situations.

2.7.1 Metabolic Cost

In looking to develop a more biologically plausible model for adaptive agents, an area of specific interest is that of metabolic cost. All of an agent's actuation, cognitive processing, and body processes come with a metabolic cost. Evolutionary pressures lead organisms towards advantageous trade-off points between sensorimotor apparatus and the metabolic cost of that apparatus, but for artificial agents it can be easy to overlook this aspect.

This is especially true for the cognitive cost, where the processing requirements of a simulation, or of the algorithms used by an agent, are easy to overlook. In nature we know that the human brain uses over 20% (Clark and Sokoloff, 1999) of consumed energy, and more recent research confirms that the majority of this energy is spent on cerebral processing (Du et al., 2008).

Empowerment has previously not accounted for metabolic cost, but the work presented in Papers 2 & 3 introduce some cognitive constraints. However, these constraints are not

intended to necessarily deal with metabolic costs, but rather with the concepts of bounded rationality (see section 2.5.2). Therefore, metabolic costs remain largely unaddressed in the present work, but are discussed in more detail in Chapter 3.

2.7.2 God View

Empowerment requires what might be called a 'God view' of the world, meaning that in order to calculate the empowerment value for a state, it is necessary to know the distribution of outcomes for every possible sequence of moves from that state.

Stochasticity can be introduced into the world in various ways, and must be represented in the channel, $p(s|a)$, on which empowerment is calculated. When that world includes stochasticity it is necessary to measure multiple samples of the outcomes for each action or action sequence.

Examples of stochasticity include the presence of any noise in the environment or the presence of other agents that can only be modelled non-deterministically (they may be acting deterministically, but with data not available to the empowerment agent).

Being able to sample the dynamics of the world in an idempotent fashion is neither plausible in a biological setting or feasible in a game scenario.

However, we note that other techniques suffer from similar problems; in looking at reinforcement learning we can see near identical shortcomings and, in fact, we could make such a criticism of almost any supervised model.

It is worth noting that empowerment does not require a channel representation that perfectly aligns to the real dynamics, and is capable of operating on imperfect representations of the world (albeit with an impact in utility). In biological scenarios we would hypothesise that evolution has equipped an organism with some domain knowledge, and in games scenarios the player knows a lot about the transition table. In both scenarios an agent can build models of the other agents based upon experience, and refine this representation.

So whilst having an internal model of the world dynamics allows an agent to use an empowerment based approach, it does mean that previously un-encountered options in an environment represent a level of risk. In order to incorporate them into a model, the agent has to test them out, which may lead to irrevocable negative situations, thus introducing a risk to exploration.

In Chapter 6, I discuss the impact of some of the work in this dissertation on improving the options for explorative behaviour in a fashion that identifies it as a risk but allows for an agent to incorporate it into any action policy.

2.7.3 Horizon

With the introduction of n-step empowerment above, we can see that empowerment introduces a 'horizon'. The n parameter specifies the number of discrete time steps into that the future the algorithm should consider when calculating empowerment.

The introduction of such a horizon, whilst drastically improving the efficacy of the empowerment approach, introduces a necessity to select the n value. It is difficult to know in advance, even with knowledge of the agent's embodiment and the dynamics of the world, what values for n may be appropriate. A key hypothesis of empowerment is that the agent's embodiment should help identify preferred states (and with the introduction of this work, actions) but the horizon parameter can have an external effect on that.

In many scenarios, be they biological or game based, it is likely that there is an appropriate or natural horizon which maximises performance. In biological scenarios this would probably be encapsulated within the embodiment, by way of the cognitive abilities of the organism.

Furthermore, it is probably that different aspects of a niche lend themselves to a different horizon, and depending on what task an agent or organism is undertaking the horizon must adapt accordingly.

In the framework of empowerment, I will highlight how different horizon values can lead to various effects such as 'boredom' or 'blindness' (Paper 1), and identify initial methods for extending the horizon value iteratively (Paper 2).

2.7.4 All States are Equal

In a discrete scenario, empowerment provides a measure of utility based upon evaluating states by measuring the observable number of new states that can be reached from each starting state within a given time horizon. However, prior to the work presented in this dissertation there was nothing built into the model that accounted for any differences in utility of the states reachable in that time.

For example, two starting states would have the same 3-step empowerment if from each an agent could reach an identical number of unique states within that 3-step horizon. However, it may be that the future states reachable from one of those starting states have no ongoing empowerment (i.e. the agent then has no ongoing moves available to them), and the other starting state reaches states that have, themselves, a higher empowerment than the starting state.

In this case empowerment would not distinguish between the two initial states, yet one has a far higher, and perhaps increasing, ongoing utility, while the other is a 'dead end'. In such a scenario, empowerment would fail to reflect this important difference, which seriously impairs it as a method for producing a sustainable action policy.

Without being addressed otherwise, it means that picking an appropriate horizon can be critical to the success of using empowerment in different scenarios.

In order to migrate from using empowerment solely for evaluating states to suggestion actions, there needed to be a method of measuring future empowerment, and incorporating it into the established framework. Proposals for such methods are introduced in Papers 2 & 3.

2.7.5 States versus Actions

In Section 2.7.4, we identified how some future states may be more empowering than others and how this current work enables empowerment to identify those, which is a critical step towards extending empowerment towards actions.

Previously empowerment could be used as a general utility measure for comparing the utility of one state with another, and in order to do this it would produce a distribution over the available actions for each state which would maximise that utility. However, that distribution of actions represented the *potential* for the agent to controllably change its current state to another, but indicated nothing about maintaining that control.

The work in Paper 2 introduces methods for differentiating future states, as discussed above, such that the action distributions become empowering, and then Paper 3 extends this further to introduce an understanding of strategies.

2.7.6 Simulation versus Reality

> *In theory there is no difference between theory and practice; in practice there is.*
>
> — *Unknown*

With many methods that developed inside simulations, where the world is well structured with controlled noise, and few peripheral features, attempts to then transfer those learnings to 'the real world' are often fraught with obstacles.

The real world is far more complex than any simulation, with variable noise (including noise in the sensors) and is continuous.

Empowerment has had limited applications to real world scenarios, and it is likely that attempts to apply empowerment in such scenarios will face many challenges related to these differences between simulation and reality. This is especially true given the relative computational complexity of empowerment.

The present dissertation does not attempt to address this issue, but I would highlight that we recognise that empowerment, being modelled by information theory, lends itself well to insight. Empowerment is designed to help guide research into this space, but (at least in its current form) is unlikely to be the 'answer' (if such a thing exists).

It is unlikely that organisms are calculating entropy, and more likely that heuristics for similar concepts are being used - therefore empowerment being a framework to investigate that inside simulations does not preclude the underlying hypothesis from being true in the real world.

Chapter 3

Structure and Preferred States

Structure and Preferred States

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

— Feynman (1965)

3.1 Structure and Preferred States

In many environments it is often easy for an observer to identify structures within that environment, whether it be natural or created; the slope on the side of a hill, the viscosity of sand at the bottom of an ocean, or the checked pattern on a chessboard. These types of environmental structure are clear.

It is well understood that over time, evolution adapts successive generations of organisms to thrive in the niche in which they exist (Mayr, 1942). There are various definitions of *niche*, in the ecological sense (Grinnell, 1917; Elton, 1927; Hutchinson, 1957). However, in each of these definitions the specifications of a niche are entirely encapsulated by the structure of the environment, the presence of other organisms within that same environment, and the behaviour of the organism.

This evolutionary process drives an organism's embodiment towards an advantageous trade-off of actuation abilities and metabolic cost, bestowing it with motor structures that provide the best chance of success at the lowest energetic cost.

These principles can be seen across the natural world. One of the earliest examples was Darwin's observations of the various finches across the Galápagos islands (Darwin, 1859); Darwin observed that the different finches had, amongst other differences, differing sizes and shapes of beak, with each species having adapted according to the available food sources.

Another example would be spiders, and their methods of hunting; only approximately half of spiders hunt using webs. Scytodidae (Spitting Spiders) also spin silk but weave no webs (perhaps because they were not effective in a new environment (Callaway, 2017)), and instead they spit the gluey substrate at their prey.

A final example, first recorded by Aristotle in *Historia Animalium* (Aristotle, 1910), and studied extensively since (Kwak and Kim, 2010) is that of gecko feet. Geckos' feet have chemical and physical structures providing powerful adhesive capabilities allowing geckos to walk on vertical surfaces and even upside down. The leg structure of the gecko also ensures that geckos can 'peel' their foot off of surfaces, so whilst they can adhere with minimal effort, thanks to their adhesive feet, they are not stuck.

Following from the observation that biological organisms are under an evolutionary pressure to modify their physiology towards the most efficient form for movement/actuation, there would be a corresponding pressure on perception. In this instance, it follows that organisms have an evolutionary pressure to adapt the physiology such that they can perceive the pertinent aspects of their environment for their survival, whilst minimising metabolic cost. Once again, this adaptation is based on the structure of their environment.

In this case, examples would include how certain species of fish living in darkness evolved away from eyes (Rétaux and Casane, 2013), and often developed alternative sensory capabilities, such as larger olfactory pits and higher chemosensory capabilities, to compensate.

3.1.1 Physiology, Preferred States and Empowerment

Having identified an evolutionary pressure on the physiology of organisms to adapt towards maximum efficiency according to the structure of their environment, it follows that we could examine the physiology of an organism and draw some conclusions about its likely habitat or niche (such study is part of ecomorphology). Biologists do this regularly when examining fossils or remains of extinct or un-encountered species (Janis and Thomason, 1995), or retroactively when studying species (Bauer et al., 1998; Kappelman, 1988).

If we return to our gecko feet example, if we were to study a gecko without previous knowledge of its habitats or behaviours, we could conclude from physiology alone that it has evolutionary reasons to be able to walk on vertical and inverted surfaces, indicating some of its preferred states.

Correspondingly, if we examine an organism in its environment, which has been equipped by evolution with the most appropriate sensorimotor apparatus, it follows that we could predict its success by measuring its potential to utilise that sensorimotor apparatus. This is exactly what empowerment measures.

Looking from this perspective we can understand that empowerment and the structure of the environment are very closely related, and we can begin to understand why it is that empowerment might demonstrate the successes that it has.

However, empowerment remains unencumbered by the metabolic costs of processing every potential future, which would be a significant cost for any biological organism or artificial agent. So, whilst empowerment seems to leverage the 'evolutionary knowledge' of the environmental structure embedded in an organism's physiology, it seems unlikely that naive empowerment is a realistic model of such organisms.

Yet, the means by which empowerment functions, by looking to maximise the potential use of their sensorimotor apparatus and thus maximise the available affordances, does seem

to be plausible; this could lead to improved understanding of these behaviours in nature, and help development of improved methods for artificial agents.

3.1.2 Highlighting Hidden Structures

In order to test the hypothesis that the structure of the world contains lower level, unseen, structures that would be accessible to algorithms like empowerment, I conducted a simple set of initial experiments to highlight this.

These experiments were based in a simple gridworld, as introduced in Klyubin et al. (2005a). In Klyubin et al. (2004a), the mutual information between an agent's chains of actuations and their starting positions relative to the centre was investigated. Automatons were produced which encoded information about an agent's starting position relative to the centre of the world, given a chain of actions (simple cardinal direction movements).

The initial experiment in Klyubin et al. (2004a) utilises mutual information, which is at the core of empowerment, and thus provided a good starting point. The experiments were replicated, to form the basis for additional experiments to investigate the structures that might be playing a role, and to begin to understand how the algorithm might be leveraging them.

The experiments took place in an infinite gridworld, within which an agent was embodied. The agent had a sensor able to return 4 values, and an actuator also with 4 values. At any time step, the sensor value would be the cardinal direction in which the agent must head in order to return to the centre of the world, akin to a gradient follower. The 4 available actuations were movement in one of these directions into an adjacent cell in the gridworld. The agent has a controller, which maps sensor values to actions: $(S_t) \mapsto (A_t)$.

In its original configuration, as described in Klyubin et al. (2004a) the mapping of sensor values to actuations performed was a simple identity mapping, to produce a gradient following behaviour. When the maximum gradient was equal in more than one direction (on a diagonal, or in the very centre), the sensor would return a random value from the available choices.

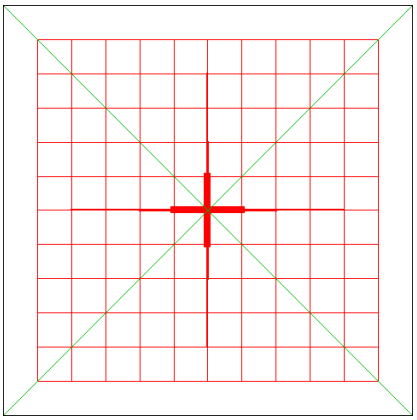


Figure 3.1: *Visual representation of structure for behaviour 27 (identity behaviour). Thicker lines show more popular paths/routes for the agent.*

To then investigate what structure of the world remained invariant of the sensor-actuator mapping, and which structures were revealed by various mappings, all possible 256 such mappings, or behaviours, were created. Each behaviour had a unique mapping of sensor value received to actuation performed, and were numbered 0 to 255. The mapping key begins with the four actions (indexed on the 4 sensor values in the order $NSEW$) set to be $NNNN$ for behaviour 0, $NNNS$ for behaviour 1, $NNSN$ for behaviour 4; always working from right to left rotating through the 4 cardinal directions and incrementing the behaviour number.

For each behaviour, the agent's initial position was set in turn to each of all 121 cells within an 11 by 11 area centred on the centre square, and was then run for 10 time steps. For all 121 runs of each behaviour, a probability distribution of the average number of transitions from one cell to another was constructed, and a visual representation was produced. Fig. 3.1 shows the original gradient follower behaviour, and it can be seen that the agent quickly reaches the centre cell and then spends the remaining time moving in and out of the centre randomly.

Even though the scenario was in a basic environment and used a limited set of possible actions and sensor values, some interesting and rich results were observed. For example, behaviour 78 is a gradient-repeller, which moves in the opposite direction given by the sensor.

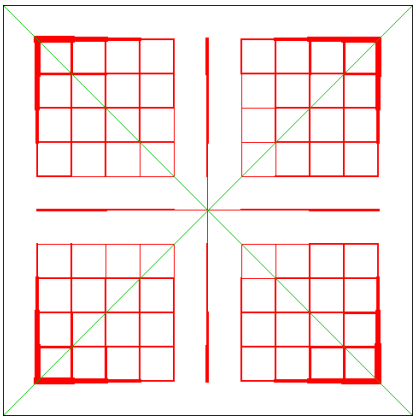


Figure 3.2: *Visual representation of structure for behaviour 78, the gradient repeller. The results clearly show the 4 primary quadrants being grouped.*

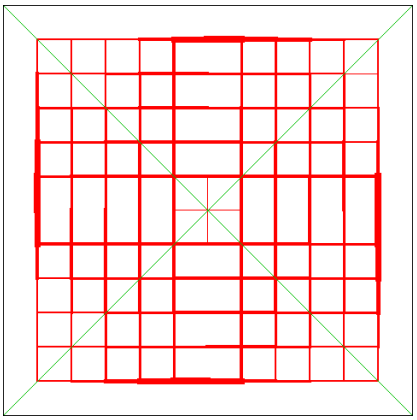


Figure 3.3: *Behaviour 180 would rotate clockwise.*

The results, shown in Fig. 3.2, were surprisingly perspicacious; the world has been neatly categorised into the 4 primary quadrants separated by the 4 cardinal directions.

A couple of other illustrative behaviours are 180 and 40. Behaviour 180 resulted in a circling behaviour, with the agent rotating continually clockwise around the source of the gradient. Behaviour 40 is an example of an asymmetric behaviour, where the sensor mapping was not a simple rotation or mirroring, and the behaviour can be seen in Fig. 3.4. The behaviour tends to the South-West corner, but when the agent starts in the North-East quadrant it always

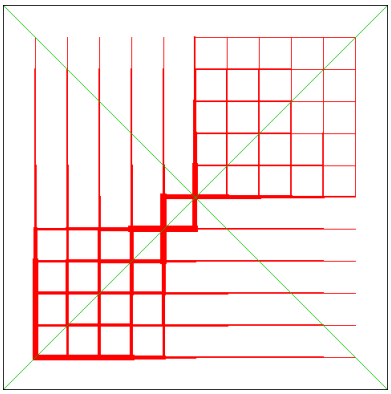


Figure 3.4: *Behaviour 40 was an asymmetric behaviour which sent all agents to the South-West. Note that agents beginning North-East of the centre always passed through the centre square - which could be thought of as a subgoal.*

travels through the centre pointer before proceeding South-West, which looks akin to a sub-goal. Important to note, regarding behaviour 40, is that the output is different to a gradient ascent to the South-West most cell, which would not display the same sub-goal behaviour.

The results of the experiment were not conclusive evidence of any particular structural observation, but successfully highlighted that a variety of structures / patterns are encoded in even simple scenarios. Further work in similar scenarios (Polani, 2011) has highlighted that agents have a higher cost to encode optimal action policies for path finding scenarios in these types of world when these structures are disrupted (i.e the cardinal directions for each specific tile are in different directions, similarly to the change in the presented scenario).

Although there is an argument that, in the experiment presented, the structure was a result of the hand-coded embodiment for our agent, the sensors and actuators we chose, it is important to realise that this type of structure (the kind the agent can detect at some low level) is precisely the type of structure that empowerment can exploit. In this case the 'concept' of quadrants emerged, and this type of concept could be possibly used advantageously were the agent to be given a task.

3.2 Paper 1: Introduction

3.2.1 Empowerment and Centrality

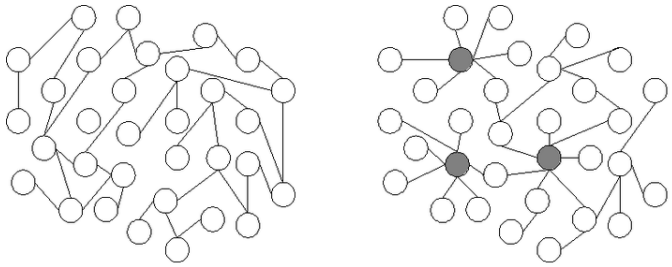
From these initial experiments, I wanted to examine further types of structure, and more directly link them with empowerment. In order to do this, and in contrast to the previous experiment, I looked to draw a parallel between empowerment and another measure of utility.

The goal of the experiments would be to further understand how empowerment could detect structure, and how such a generalised mechanism would compare to something that was hand crafted to explicitly examine the structure, both in terms of how the structure was detected, and whether the two measures would see any correlation.

I believed it was best to select a utility that was designed to work in highly structured environments, and that was hand crafted to measure some element of fitness or utility in that environment.

The domain I identified as having good properties was graph theory, which has been researched in depth and provides a rigorous set of tools with which to measure structure and utility in a variety of ways. Furthermore, graphs can be used to relatively neatly represent the state space of deterministic scenarios, with the vertices representing a given state and the edges between them representing actions that would change the state to that represented by the destination vertex. In simple geographical scenarios where the state of the world is represented by nothing more than the position of a single agent, such as that represented by the gridworld scenarios used above, this representation also works very neatly as a graph representation can be conveniently overlaid onto the geographical representation.

Graphs can take on a variety of forms, with their degree distribution and topologies allowing them to fall into a variety of categories based on their structure. One very interesting case is that of scale-free networks, in which the degree distribution follows a power law, in which there are a few vertices with a high degree, but most vertices have a far lower degree. Their



Random Graph Scale-Free Graph

Figure 3.5: *Examples of random and scale-free graphs, each with 32 nodes (Castillo, 2004).*

typical structure is independent of the graph's size; with fewer or more vertices, the graph would still exhibit similar properties. The exact distribution of edges per vertices follows a power law distribution (Barabási and Albert, 1999):

$$P(k) \sim k^{-\gamma}. \tag{3.1}$$

where $P(k)$ is the probability that a vertex connects with k other vertices, and decreases exponentially according to the coefficient γ.

In recent years scale-free graphs have received a lot of attention since it was identified that the network formed by the world wide web was such a graph (Barabási and Albert, 1999). Scale-free graphs can also be seen in many real world situations, including protein interaction networks (Jeong et al., 2001), and social networks (Barabási, 2003).

Scale-free graphs provided an abundance of structure, and also neatly indicated a measure which could be used as a comparison to empowerment: centrality. The group of measures known as centrality are well established measures, used to measure the relative importances of different vertices in a given graph. There are various centrality measures which suit different scenarios and which measure importance in varying ways.

The measure that most closely resembles empowerment is *closeness centrality* (Bavelas, 1950), which for any vertex in a graph measures the average distance from that vertex to

all other vertices in the graph. The measure is commonly inverted, so that higher values represented an improved level of centrality.

Closeness centrality is similar in spirit to n-step empowerment; both measure how many new states you can reach from your current state, and can be compared to the same measure at other states in the space. However, one fundamental difference between the two measures is that closeness centrality is a global measure, that takes the complete graph into account, whereas n-step empowerment has a horizon which may be large enough to encapsulate the total state space, but usually is not.

This apparent disparity between a global measure and a local measure became an interesting aspect of the results, following two experiments comparing the measures.

3.2.2 Experimental Setup

In Paper 1, I conducted the initial experiments by creating a graph that represented the familiar 2D gridworld scenario; in this instance a box was added to the scenario. In any given time-step the agent moved as before, by either staying still or by moving by a single cell in any of the four cardinal directions. This scenario was first introduced in an empowerment context in Klyubin et al. (2005a).

Should the agent be in a cell adjacent to the box, and move towards the box then the box would move one cell in the same direction and the agent would move into the cell that was occupied by the box.

A graph representation of the scenario was created, in which each vertex represented a state of the world, given by the position of the agent and the box, and the edges between the vertices representing an action by the agent that moves the agent and possibly the box.

In the original experiment the world was infinite, and the agent 'preferred' (as indicated by higher empowerment) to be near the box, which provided an affordance and thus increased empowerment by allowing the agent to potentially reach a wider variety of future states.

This setup allowed me to use established results for empowerment, and apply closeness centrality to the same problem to draw direct comparisons between the two utilities. I conducted additional experiments that utilised the scale-free property of graphs as a second comparison between an established and understood results (in this instance based on the non-empowerment utility).

Paper 1 was published in *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems* and presented at the *Artificial Life XI* conference (Anthony et al., 2008).

On Preferred States of Agents: how Global Structure is reflected in Local Structure

Tom Anthony[1], Daniel Polani[1,2] and Chrystopher L. Nehaniv[1,2]

[1]Adaptive Systems Research Group
[2]Algorithms Research Group
School of Computer Science, University of Hertfordshire
College Lane, Hatfield, Herts, AL10 9AB, UK
{T.C.Anthony,D.Polani,C.L.Nehaniv}@herts.ac.uk

Abstract

We investigate the correlation between the information theoretic measure of *empowerment* and the graph theoretic measure of *closeness centrality*, to better understand the structural conditions that must exist in a world for learning and adaptation. We examine both measures in both a simple grid-world scenario, represented as a graph, and on a scale-free graph. We show a strong correlation between the two measures, and discuss the strengths and weaknesses of both. We go on to show how the local measurement of empowerment can in many cases predict a measure for the global measurement of closeness centrality.

Motivation

"Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry." - Richard Feynman

Learning and adaptation are central themes to artificial life, and it is our hypothesis that a better understanding of the conditions that must exist to make learning, adaptation and evolution possible will help to guide future research. It is plausible to assume that an arbitrary or random world would be extremely difficult, if at all possible, to learn. We know there is significant structure in the world, and believe that learning takes advantage of this structure. In this paper we begin to investigate what conditions, embedded within a world through some underlying structure, are necessary for certain types of adaptation problems.

It has been hypothesised that embodied agents receive an adaptive and evolutionary advantage by optimising their sensoric and neural configurations for their environment. Specific attention has been paid to processing and optimising of Shannon-type information they receive from their environment (Attneave, 1954; Barlow, 1959, 2001; Atick, 1992). Similar work includes the concept of *homeokinesis*, proposed by Der et al. (1999), where a homeokinetic system, or agent, learns to improve the predictive capabilities of its future perceptions.

A specific flavour of this view suggests that such informational predictive princiles could provide organisms/agents with intrisic motivation. Examples include that by Prokopenko et al. (2006) and Bialek et al. (2001), which use similar approaches based on *excess entropy / predictive information*.

In this paper we have chosen to use *empowerment* (Klyubin et al., 2005b,a), an information theoretic measure for the efficiency of a *perception-action loop*. Essentially empowerment uses the channel capacity for the external aspects of a perception-action loop to identify areas that are advantageous for an agent embodied within an environment.

It assumes areas with a high efficiency of the perception-action loop should be favoured by an agent. Based entirely on the sensors and actuators of an agent, empowerment encapsulates an evolutionary perspective; namely that evolution has selected which sensors and actuators a successful agent should have, which in turn suggests which parts of the world should be visited.

This hypothesis was tested in a variety of different scenarios (Klyubin et al., 2005b,a; Capdepuy et al., 2007), and notwithstanding the quite different scenarios it coincided surprisingly well with an intuitive understanding of favourable behaviours or of natural solutions to particular challenges of adaptation. Furthermore, it correlated well in some scenarios that had been hand crafted to evaluate the results.

Notwithstanding the successful performance, we do not currently have a strong understanding of why this may be. What are the properties of the world that make empowerment such a universal measure. Why should it work at all? This is the question we are going to study in this paper.

Locating Structure

We hypothesise that an agent that optimises its sensorimotor apparatus improves its ability to detect the underlying structure of the world, and that this is an important aspect of such optimisation. We further hypothesise that a better understanding of this structure would improve such optimisation, and thus allow for better adaptation and learning.

To investigate this we set out to start identifying the basic properties of the world, and how they are detected by empowerment. We selected to go about this by investigating a representation for an environment that manifests its structure in an easily observable manner, is well understood, and has established methods for measuring preferable states.

We chose to represent the state space using graphs, which fit all these criteria; they are well understood through graph theory and social network analysis, and they have accessible methods for identifying certain aspects of their structure. As a measure to identify preferred states we chose to use *centrality*, a measure of a node's importance from graph theory, which is a well established method (Wasserman and Faust, 1994). There are varying measures for centrality; in this paper we use *closeness centrality*, which most closely corresponds with the spirit of empowerment.

Most stationary worlds, containing an embodied agent, can be viewed of as the current state of the world connected to neighbouring states by the actions the agent would need to take to arrive at them; this can be modelled as a graph. This same representation of the world was used by Şimşek and Barto (2007) in investigating skill development among agents.

Preferred states

Clearly, for two measures to be able to correlate they must be measuring a similar property, and a quick overview of the measures should convince the reader.

Empowerment, a local measure, quantifies the changes that an embodied agent can make on its environment, and observe the effects of, in a given time period. In reducing ourselves to a simple representation of the world which is entirely deterministic, we have created a special case for empowerment. However, it can work in both entirely deterministic and probabilistic environments, which may even be non-stationary (Capdepuy et al., 2007).

The closeness centrality of a node in a graph is calculated by adding the distance of the shortest paths from that node to every other node in the network, and then inverting this value so that a shorter total path to all other nodes has a higher value. To calculate the closeness centrality of a signal node requires viewing the whole graph; it is a global measure. Klyubin et al. (2005a) showed an example where a similar measure, the average shortest distance in the case of mazes, correlated with empowerment.

We will examine two scenarios, and will employ both empowerment and closeness centrality in each for identifying and measuring states that an embodied agent would find 'interesting' or 'preferential' to be in. When we use the word 'state' we refer to the state of the whole system, including both the environment and the agent.

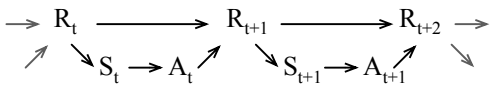


Figure 1: Bayesian network representation of the perception-action loop.

Information Theory

The notion of empowerment is based on information theory, introduced by Shannon (1948). To introduce this, the first important measure is *entropy*, which is a measure of uncertainty:

$$H(X) = -\sum p(x) \log p(x). \tag{1}$$

Where X is a discrete random variable with values $x \in X$ and $p(x)$ is the probability mass function such that $p(x) = Pr\{X = x\}$. The logarithm can be taken to any chosen base; in our paper we consistently use 2, and accordingly the units of measurement are then called *bits*. If Y is another random variable jointly distributed with X the *conditional entropy* is:

$$H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x). \tag{2}$$

This measures the remaining uncertainty about the value of Y, if we know the value of X. Finally, this also allows us to measure the *mutual information* between to random variables:

$$I(X;Y) = H(Y) - H(Y|X). \tag{3}$$

Mutual information can be thought of as the reduction in uncertainty about the variable X or Y, given that we know the value of the other. The mutual information is symmetric, so we could also use $I(X;Y) = H(X) - H(X|Y)$ (Cover and Thomas, 1991).

Empowerment

Empowerment is based on the information theoretic perception-action loop formalism introduced by Klyubin et al. (2005a, 2004), as a way to model embodied agents and their environments. The model views the world as a communication channel; when the agent performs an action, it is injecting Shannon information into the environment, which may or may not be modified, and subsequently the agent reacquires part of this information from the environment via its sensors.

In Fig.1 we can see the perception-action loop represented by a Bayesian network, where the random variable R_t represents the state of the environment, S_t the state of the sensors, and A_t the actuation selected by the agent at time t. It can be

seen that R_{t+1} depends only on the state of the environment at time t, and the action just carried out by the agent.

By modelling this as a communication channel, we can employ information-theoretic methods, which are the basis for empowerment. First, we must introduce channel capacity (Shannon, 1948; Cover and Thomas, 1991) for a discrete memoryless channel:

$$C(p(y|x)) = \max_{p(x)} I(X;Y). \tag{4}$$

The random variable X represents the distribution of messages being sent over the channel, and Y the distribution of received signals. Clearly, the higher the mutual information between the two variables, the higher the capacity of the channel. The channel capacity is measured as the maximum mutual information taken over all possible input distributions, $p(x)$, and depends only on $p(y|x)$, which is fixed. One algorithm that can be used to find this maximum is the iterative Blahut-Arimoto algorithm (Blahut, 1972).

Empowerment can be intuitively thought of as a measure of how many observable adjustments an embodied agent can make to his environment, either immediately, or in the case of n-step empowerment, over a given period of time. An alternative way to view empowerment is that it guides agents to places in the world where they get the most benefit from their sensors and actuators. Using the above perception-action loop formalism and the Blahut-Arimoto algorithm, this can be directly quantified. We remind the reader that sensors and actuators implicitly encode evolutionary knowledge of the type of information to perceive and 'create'.

In the case of n-step empowerment, we first construct a compound random variable of the last n actuations, labelled A_t^n. We now need to maximise the mutual information between this variable and the sensor readings at time $t+n$, represented by S_{t+n}. Here we consider empowerment as the channel capacity between these:

$$\mathfrak{E} = C(p(s_{t+n}|a_t^n)) = \max_{p(a_t^n)} I(A_t^n; S_{t+n}). \tag{5}$$

An agent that maximises its empowerment will position itself in the environment in a way as to maximise its options for influencing its relationship with the environment (Klyubin et al., 2005a).

Note that in this paper we are use empowerment in a deterministic scenario, within a discrete world, but that empowerment is defined in full generality for non-deterministic probabilistic environments and does not assume perfect information.

In this paper we can use a shorthand method for calculating empowerment; we are able to do this for several reasons. All the scenarios we examine are deterministic and feature no non-stationary elements, and so do not require the probabilistic elements of empowerment. Additionally, as they are all represented as a graph, we are able to further simplify the formula. We can calculate n-step empowerment for a node v_i on the graph thus:

$$\mathfrak{E}_n(v_i) = \log \left[\sum_{\substack{j=1 \\ d(v_i,v_j)\leq n}}^{g} 1 \right] \tag{6}$$

Where $d(v_i, v_j)$ is the geodesic distance between the nodes v_i and v_j. Note that this is a shorthand method we are able to use as we have complete knowledge of the scenarios and the representation; without such knowledge Eq. (5) would work just as well in the same scenarios, using the perception-action loop formalism.

Closeness Centrality

Graph Theory and Network Analysis have long had a requirement for identifying important nodes in a graph (Wasserman and Faust, 1994). The simplest methods for this have been to count the edges leaving or entering a node, known as outdegree and indegree respectively. This is very simplistic and is normally inadequate for complex graphs. Therefore, the primary method for measuring node importance is a group of various measures collectively known as centrality. There have been several methods of centrality suggested over time, but one of the most popular is closeness centrality, which can be presented in various ways. As mentioned in Wasserman and Faust (1994), and reviewed by Freeman (1979), the simplest formula for closeness centrality is that suggested by Sabidussi (1966):

$$C_C(v_i) = \left[\sum_{\substack{j=1 \\ j\neq i}}^{g} d(v_i, v_j) \right]^{-1}. \tag{7}$$

For a given node v_i, in a graph with g nodes, this gives a measurement of the sum of the shortest paths to all other nodes, which is then inverted to give a higher centrality to those with shorter total paths to the rest of the graph. Intuitively, this can be closely linked to the average distance from all other cells that empowerment was anti-correlated with, from the maze scenario used in Klyubin et al. (2005a).

To calculate the closeness centrality on the graphs encountered throughout this paper, we used the network analysis software Pajek (Batagelj and Mrvar, 1998). Pajek uses a modified version of closeness centrality, suggested in Beauchamp (1965):

$$C'_C(v_i) = \frac{(g-1)}{\left[\sum_{j=1}^{g} d(v_i, v_j)\right]} = (g-1)C_C(v_i). \tag{8}$$

This formula is used simply to normalise the closeness centrality figures to the graphs size in order to allow comparison of the figures between graphs of different sizes.

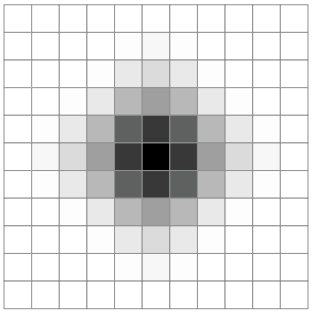


Figure 2: View of the empowerment distribution for the gridworld scenario, with the box positioned at the center. A darker shade means higher empowerment. Empowerment scales from 5.92 to 7.79 bits.

Scenarios

In order to compare these two measurements we apply them to the same two agent scenarios to identify the correlation between them, and any areas of disparity. In order to construct the first scenario, it is necessary to observe that most state spaces encompassing an agent in a stationary world can be naturally represented as a graph of nodes; each node/vertex represents a specific state in the world and the edges between them corresponding to the actions of an agent that change the state of the world.

Box pushing

Consider the box pushing scenario from Klyubin et al. (2005a) as a graph. The scenario consists of a gridworld of infinite size, within which there exists an agent and a box, each of which occupy a single cell. The box is visible to the agent; his view of the world consists of his position and the position of the box (and given the world is infinite, the relative position of the box and agent defines the state of the world). The agent has 5 actions available to it at any time; it can stand still, or move to one of the four neighbouring cells. If the agent moves into a cell that is occupied by the box then the box is pushed, in the same direction, into the adjacent cell.

In Klyubin et al. (2005a) it was shown that for any n-step empowerment, the agent prefers being near the box, which gives it more influence on the state of the world. It most 'enjoyed' beginning on top of the box, where moving in and of the 4 directions would allow it to fall down next to the box, from where it could start pushing it like normal; this could be used as a starting position but was a position impossible for it to return to.

In translating this world into a graph representation, we needed to limit our originally infinite world to a finite graph. We investigate the influence of this finiteness by examining the growth of centrality. We show that beyond a certain horizon it can be seen that the centrality increases in a continuous fashion and that the centrality for the nodes represented in previous approximations grows proportionately. Whilst we do not offer a proof of this fact, in Fig.3 we demonstrate the point by showing the correlation between graph representations of increasing diameters.

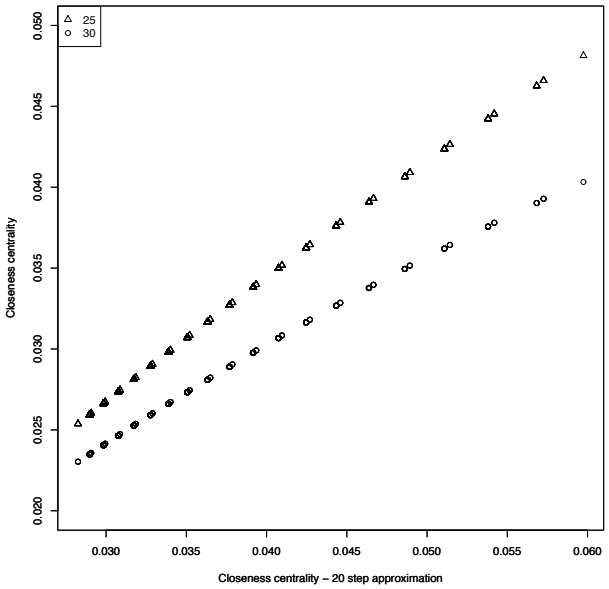


Figure 3: Correlation of closeness centrality for 25-step and 30-step graph approximations against a 20-step approximation. Each point on the graph represents a different starting state of the world (relative position of the agent to the box).

Results

Klyubin et al. (2005a) had previously shown how empowerment worked in the box pushing gridworld experiment, and so it made for a good environment in which to run our initial experiments. We generated a unweighted directed graph to represent the world. Note that we are using a non-classical view of graphs; rather than viewing them as comprised of units, with connecting links between them, we are viewing each node as a possible state of the world, including the agent itself, (of which, only one can be the real state at any moment) and the edges as transitions between these states.

To do this, we initialised the world with the box in the center, and the agent standing upon the box, as described earlier. We then let the agent run through every possible trajectory of 30 actuations, generating a graph of states and actions; the final graph had 419,121 nodes. Using Pajek, we calculated the closeness centrality for all nodes in the graph.

We next measured empowerment for every state with the box positioned in the center of the world, and the agent positioned at each location that it could reach within 30 timesteps from the center. This was sufficient as the dy-

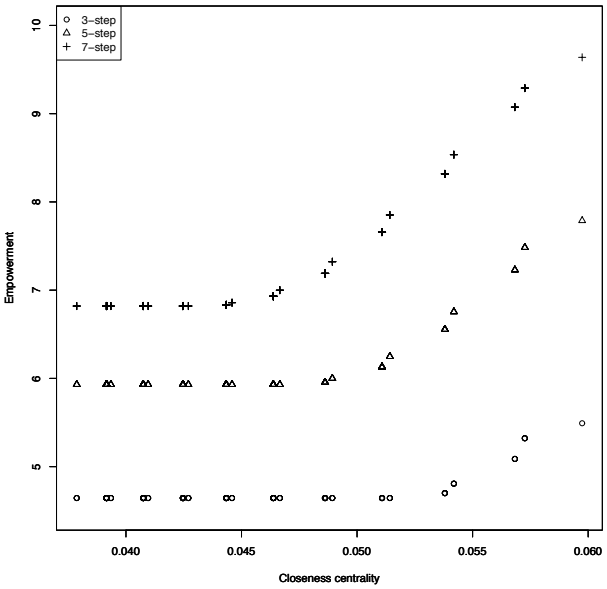


Figure 4: Correlation plot between Empowerment and Closeness Centrality (20-step graph). The horizon effect of empowerment can be seen clearly.

namics of the world comes from the agent's initial position relative to the box, and thus moving the box was unnecessary. Our empowerment measurements were run to measure 3-step, 5-step and 7-step empowerment.

In order to correlate empowerment and centrality, we collated the results, removing the centrality results for nodes where the box was not positioned in the center of the world; this gave us a state for state comparison of each measure against the other for different initial positions of the agent.

We additionally ran the same experiment for graphs produced for both 20 and 25 timesteps, to identify the influence of representing the infinite gridworld as a finite graph did not skew the results. We found that the correlation of centrality for the overlapping nodes of these varying size graphs indicates a close to linear relationship and finite graphs work as a good approximation.

Note that closeness centrality is a global property, calculated it for any given node requires seeing all other nodes in the graph, while empowerment is local and looks only at neighbouring nodes within a given distance.

Local Structure

As hypothesised, we found a very strong correlation between the closeness centrality and empowerment, which can be seen in Fig.4. The graph shows clearly the horizon effect of empowerment; it can be seen to be constant whenever the box is outside of the agent's reach. For n-step value with larger values of n the horizon can be seen to extend further from the box. Once the box is within it's reach, according to n, the empowerment grows as the agent increases its influence over the world by getting closer to the box.

The horizon effect emphasises that empowerment is a local measure; it cannot see the whole world. However, when the agent is within an area where it can improve it's ability to manipulate the state of the world, this local measure correlates with the global measure of the world given by closeness centrality.

This highlights that in an infinite, or an unexplored, world where centrality cannot be employed, empowerment provides a measure that can be used. Whilst empowerment is limited by the horizon effect, exploring the world (which would be necessary to use closeness centrality) would allow our agent to also overcome the horizon.

In addition, this correlation also confirms our hypothesis that empowerment, within its horizon, does see global aspects of a system at a local level within this world. What structure or prerequisites that must exist for this effect to take place are yet to be determined.

It is important to note that the results from empowerment can be computed by the formula in Eq. (6), or equally by that in Eq. (5), without modelling the world as a graph at all.

Scale-free Graphs

The second scenario uses scale-free networks (graphs); a very important subclass of graphs, in which there are a few nodes with a high degree, and most nodes have a far lower degree. Their typical structure is independent of the graph's size; with fewer or more nodes, the graph would still exhibit similar properties. The exact distribution of edges per node follows a power law distribution (Barabasi and Albert, 1999):

$$P(k) \sim k^{-\gamma}. \tag{9}$$

Here $P(k)$ is the probability that a node connects with k other nodes, and decreases exponentially according to the coefficient γ.

As discussed in Barabasi (2003), scale-free graphs can be seen in many real world situations, including protein interaction networks (Jeong et al., 2001), social networks, and even the world wide web (Barabasi and Albert, 1999).

We hypothesise that the scale-free property of graphs can work to synthesise an underlying structure that may be found in real world task spaces, and can be used as a good platform for initial investigation of such structure.

Results

Using preferential attachment algorithm introduced by Barabasi and Albert (1999) we constructed a scale-free undirected graph with 400,000 nodes to run our measures on. Our graph was built using an initial complete graph of 3 nodes, and adding additional nodes one at a time. Each

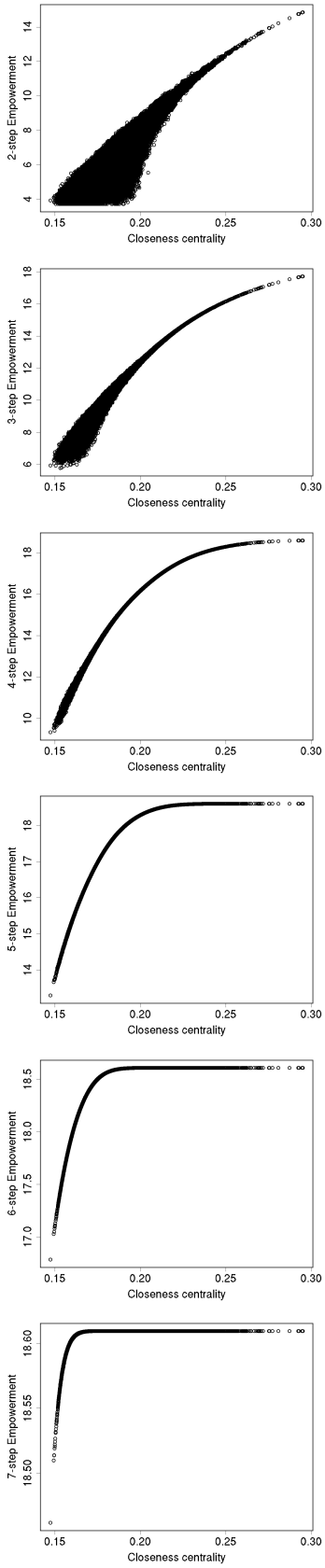


Figure 5: Correlation between closeness centrality and 2-step to 7-step empowerment.

new node would create 3 new edges connected to 3 different nodes on the existing graph, chosen using a probability according to their current degree.

For all nodes in the graph we calculate both the n-step empowerment (for a range of values of n) and the closeness centrality. To calculate the closeness centrality, we again use the Pajek analysis software.

Our results here corroborate those from our first experiment with regard to the correlation between closeness centrality and empowerment. Here, we see the inverse of the horizon effect; given too much time, empowerment can reach any part of the graph (analogous to being able to do anything within a world) and assigns almost all nodes equal value. This is an interesting point for empowerment; given too high of a 'budget', where an agent can do everything possible within the world (or reach every node in a graph) then it does not differentiate between them. This is the type of world we would describe as 'boring'; one where and agent can do anything it wants from any position of the world.

Again though, empowerment sees at a local level aspects of the global property of the world. In this scenario, this is maybe not surprising given the nature of a scale-free graph; but it is important to see that empowerment was not told anything of the structure of the world, and that still this fact comes through.

In Fig.5 we show the correlation between closeness centrality and n-step empowerment for n=2 to n=7. Note that even 2-step empowerment has a strong correlation at the higher centrality nodes, and 3-step even more so. As n increases it can be seen that the small-world property of the graph results in an empowerment ceiling being reached which results in a reduced correlation for high centrality nodes.

Discussion

Both of our experiments highlight the strong correlation between empowerment and closeness centrality, and that even n-step empowerment with a low value for n will normally serve a a strong predictor for centrality. This is significant given that individual node centrality is a global property of a graph, but we can use a local measure to give similar relative values to nodes. Note that empowerment doesn't see any more than centrality, but in the 'interesting' parts of the world it does see, the two measures agree.

In both scenarios the correlation is strong provided that the n chosen for n-step empowerment is suitable. We believe a simple method for overcoming this in an unknown world is for an agent to select the lowest n value possible; if the horizon of this n does not allow the agent to observe any degrees of freedom it can then increase n incrementally to overcome this (or embark on a random exploration).

With empowerment, selection of a suitable n is interesting in another regard; a low value of n can mean encounter-

ing the horizon effect, and possibly not seeing 'interesting' parts of the world, whilst a too high value of n can result in the agent being able to do anything and not needing to distinguish between different states. The result of this is a particular world having an n value with the correct balance between these two effects, which we hypothesise may reflect one aspect of the underlying structure which is important for learning and adaptation.

Closeness centrality is limited to deterministic task spaces that can be completely represented by either a directed or undirected graph, which constricts the space of problems it can be used to measure. In the space of problems in which both measures can be used, these results indicate not only empowerment correlates well with centrality, but it does so without complete knowledge of the world. Furthermore, it can work in non-deterministic, non-stationary, environments which cannot be represented as a graph, including infinite worlds.

A comparison could have been drawn between the global measure of closeness centrality, and some local version of centrality that worked on a local subset of the graph, and probably a similar correlation would have been found. However, any such localised version of centrality would suffer from many of the same restrictions that centrality does compared to empowerment. Empowerment has been chosen as part of a much more general picture, and as well as including an evolutionary aspect, will allow us to extend the research into non-deterministic environments in future work. Essentially we are using centrality as a 'sanity check' that empowerment does something sensible in these scenarios.

Overall, we believe that these results show a strong indication of certain global aspects of various worlds being 'coded' at a local level, and an appropriate sensory configuration can not only detect this information, but can also use it. Such uses could include learning and adaptation, and uses for evolution between generations. There are indications that understanding which aspects of global structure are visible at a local level would allow improved adaptation and learning for agents embodied within the corresponding world.

Future work

Vergassola et al. (2007) drew a parallel between the behaviour of biological organisms and search methods that use local informational cues to draw conclusions about the global structure of the world. It is our belief that further study of this area will allow us to not only draw further parallels with the learning and adaptation methods employed by biological organisms, but will also allow a better understanding of these processes leading to improved methods.

Further work needs to be done to extend these results into other worlds and task spaces, and to better understand in which scenarios they hold true. This should include worlds with various elements providing opportunities for agents to manipulate their environment, and even non-stationary worlds.

Attention needs to be paid to how to choose an initial strategy when presented with a completely unknown task space (such as choosing an initial n for empowerment) and conversely, how much of this information is embedded with an agent or organisms embodiment.

Acknowledgements

The authors would like to thank the reviewers for their feedback and suggestions which helped bring about various improvements to the paper.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems*, 3(2):213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.

Barabasi, A.-L. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume Books.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barlow, H. B. (1959). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217–234. The M.I.T. Press.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253.

Batagelj, V. and Mrvar, A. (1998). Pajek – program for large network analysis.

Beauchamp, M. A. (1965). An improved index of centrality. *Behavioral Science*, 2:161–163.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comp.*, 13(11):2409–2463.

Blahut, R. (1972). Computation of channel capacity and rate distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473.

Capdepuy, P., Polani, D., and Nehaniv, C. L. (2007). Maximization of potential information flow as a universal utility for collective behaviour. In *Proceedings of the First IEEE Symposium on Artificial Life*.

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

Şimşek, O. and Barto, A. (2007). Betweenness centrality as a basis for forming skills. Technical Report TR-2007-26, University of Massachusetts, Department of Computer Science.

Der, R., Steinmetz, U., and Pasemann, F. (1999). Homeokinesis - a new principle to back up evolution with learning. In Mohammadian, M., editor, *Computational Intelligence for Modelling, Control, and Automation*, volume 55 of *Concurrent Systems Engineering Series*, pages 43–47. IOS Press.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

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

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005a). All else being equal be empowered. In Capcarrère, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life: Proceedings of the 8th European Conference on Artificial Life*, volume 3630 of *Lecture Notes in Artificial Intelligence*, pages 744–753. Springer.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005b). Empowerment: A universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE Press.

Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). Evolving spatiotemporal coordination in a modular robotic system. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., and Parisi, D., editors, *From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy, September 25-29 2006*, volume 4095 of *Lecture Notes in Computer Science*, pages 558–569. Springer.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Vergassola, M., Villermaux, E., and Shraiman, B. I. (2007). 'infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis*. Cambridge University Press.

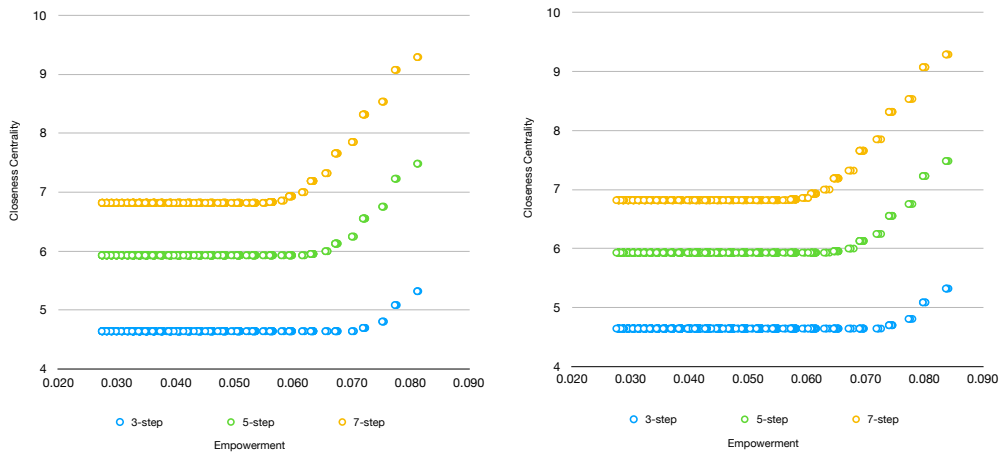


Figure 3.6: *Correlation plot between Empowerment and Closeness Centrality. On the left for the agent starting in position 3, 3 (relative to the central box), and on the right for the agent starting in 6, 1.*

3.4 Paper 1: Supplemental Result

In the paper, a box pushing scenario was presented in which a correlation between empowerment and closeness centrality was demonstrated. The scenario compared multiple relative starting positions of the agent to the box for multiple empowerment horizons against closeness centrality for the same position. The graph for closeness centrality was generated from the default position of having the agent stood atop of the box, and allowing the agent to run for 20 time steps; the states discovered were then used as the starting position for three empowerment runs (for 3, 5 and 7 time steps).

Because what matters is the relative position of the agent to the box (as the world away from the box is homogenous), it seems intuitive that generating the graph from an alternative default state should not have change the nature of the correlation between the two metrics. However, in order to confirm that this is indeed true, and to add clarity to the result as presented in the paper, I ran the experiment again from different initial configurations to verify.

In Fig. 3.6 we can see results for two additional starting configurations, demonstrating that the centrality range shifts but the the pattern of correlation remains.

3.5 Paper 1: Discussion

3.5.1 Horizon and Local Structure

The paper discusses the challenges of selecting a suitable value of n for the number of steps available to empowerment. This 'horizon' is an interesting aspect of empowerment for several reasons.

Perceptual horizon vs Temporal Horizon

Note that the scenarios presented in the paper have discrete time steps and a discrete perceptual boundary. Such a model is common for game playing scenarios, but for biological scenarios both of these would be continuous.

The scenarios are based on previous experiments and the 'horizon' discussed simply refers to the temporal horizon, defined by the value of n. We recognise that the optimal value for n (for the success of an agent in a survival scenario) must account for the embodiment of the agent, which includes an implicit understanding of its perceptual boundaries, which indicates likely trade-offs between these two values.

This thesis does not attempt to reconcile the potential interplay between these two facets; it may be that there is an interesting set of experiments to do to understand how the value n changes given improved or reduced perceptual abilities for an agent. This question would also play to the metabolic cost of those capabilities (both perceptual and cognitive with regards to planning temporally).

Global vs Local Structure

The paper discusses the relationship between a 'global' (centrality) metric and a 'local' one (empowerment), and discusses how empowerment can correlate with the global metric within the range of empowerment's horizon value.

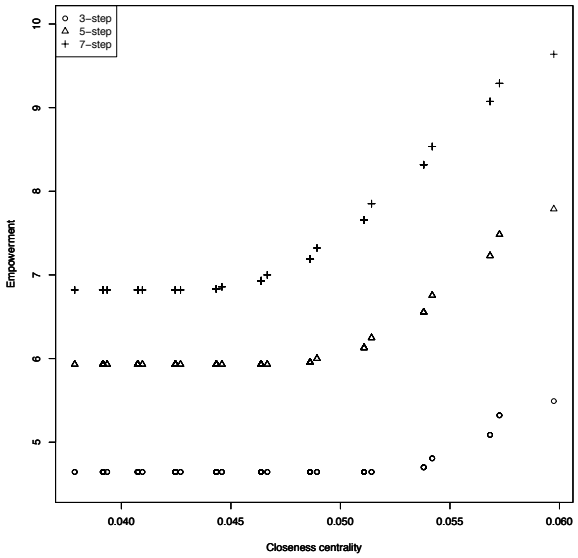
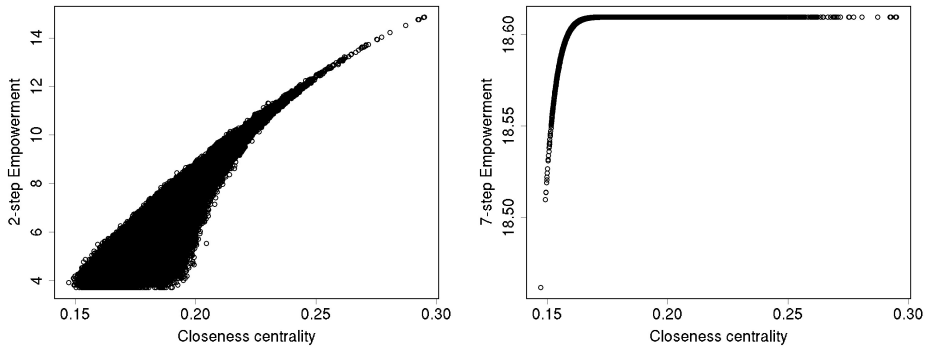


Figure 3.7: *Correlation plot between Empowerment and Closeness Centrality. The horizon effect of empowerment can be seen clearly.*

Returning to Fig. 3.7, we can see that as the centrality value increases we reach a point where empowerment begins to increase relative to the increase in centrality. The point where this correlation between centrality and empowerment begins to take effect is a function of the horizon length of empowerment, and there is a hard cut-off point. We can see that with 7-step empowerment, the correlation begins the soonest, and in every case the correlation continues for the remainder of the vertices in the graph.

This correlation between local measurements and global measurements would break down in larger, more complex scenarios where there were a variety of features that provided affordances to the agent. If we imagine a similar gridworld, but one in which there are dozens of boxes spaced out equally in the centre of the world, then it is immediately intuitive that the position of highest centrality would be that in which the agent started in the centre of the boxes. However, for empowerment, we might see many small peaks of utility, all deemed to have equal empowerment, where the agent is positioned such that within his time horizon he can only manipulate the closest box.



Blindness Boredom

Figure 3.8: *On the left we can see that 2-step empowerment has 'blindness' for low-centrality values, whereas the right shows that 7-step empowerment suffers from 'boredom' for high-centrality values. These are two types of inability to differentiate states.*

Such a scenario highlights the trade-off that evolution must face in deciding what sensory physiology an organism requires vs the metabolic cost. Being able to select a position, with certainty, that is the exact centre may not be worth the metabolic cost of the sensory apparatus required to do so.

The important observation is that when local structure gives an indication into global structure, empowerment is capable of 'detecting' it and using it. Its capability to detect such structure is entirely a function of the horizon.

Boredom vs Blindness

In the previous section, we have discussed the problem of potential 'blindness' when the empowerment horizon is too short to detect certain structures in the environment. The opposite of this is horizon values that lead to 'getting bored', or seeing no differences anymore. These are both a form of being unable to differentiate states from one another. The former (discussed in the previous section) underestimates the utility of some states, and the latter overestimates the relative utility of states.

This problem of 'boredom' demonstrates that, even if we could ignore issues around metabolic, cognitive or processing cost, simply picking the longest possible horizon for empowerment would not overcome the issue of state differentiation.

3.5.2 Generality

We have seen that empowerment correlates well with hand-picked metrics that are designed for the scenario. Whilst correlation is perhaps not surprising retrospectively, what is important is that empowerment is totally general, and 'knows' nothing about these scenarios.

The empowerment models for the paper could be represented as graph, but empowerment was calculated purely on the state transitions, and thus were not in any fashion tuned towards a graph scenario rather than some other model.

Furthermore, we note that many scenarios could not be transformed into a graph representation and so centrality could not be applied to such a scenario. The most obvious class of scenario is anything non-deterministic where the transitions between states are in some way stochastic.

Thus we see that empowerment is entirely general and can apply to complex scenarios without needing domain specific knowledge, and can deal with non-deterministic situations. In looking to build upon the work of this paper, it was important that this aspect of empowerment remained intact.

3.5.3 Conclusions and Direction

Following this paper, there were several aspects that I concluded needed to be addressed in order to leverage the strengths of empowerment for a wider range of scenarios, based upon my motivation to develop a universal utility which was capable of helping to drive intelligent behaviours:

- Address the issue of a hard 'cut off' beyond which empowerment cannot differentiate states. Selecting the 'perfect' horizon is not only going to be probably intractable in many environments, but is also unlikely to be static across states in the same environment, and should also be capable of being adaptive depending on available time/resource.
- Certain states appear identical in terms of utility, but the hard horizon applies equally to the reachable future states that contribute to the calculation of that empowerment. The presence of this hard horizon means that any sort of utility hill-climbing is not viable.
- In the conclusion of the paper, I touched on the need to better understand how to 'choose an initial strategy when presented with a completely unknown task space'. This fitted with my broader motivations and meant moving from evaluating only states to also being able to not only evaluate actions, but pick sets of actions in a self-motivated manner which could drive behaviours.

Chapter 4

Action Selection: Self-Motivation

Action Selection: Self-Motivation

Knowing is not enough; we must apply. Being willing is not enough; we must do.

— von Goethe (1870)

4.1 Empowerment and Compression

An equivalency between compression and intelligence has been suggested in various forms in the field of Artificial Intelligence (Hutter, 2005, 2001; Mahoney, 1999; Schmidhuber, 1992; Dowe et al., 2011; Dowe and Hajek, 1997; Chaitin, 1982; Hernández-Orallo and Minaya-Collado, 1998) for many years; since 2006 the Hutter Prize has been running, offering $50,000 for improvements in lossless natural language compression directly motivated by this belief. Opinion is divided on how lossless or lossy compression affects improved intelligence, but the competition organisers acknowledge human-like compression is lossy (Hutter, 2006; Mahoney, 2006); it seems clear that compression of some sort is necessary for intelligence to emerge. It is not clear whether there are restrictions on how that compression is manifested, and various solutions have been proposed (Hutter, 2001; Dowe et al., 2011); of particular interest to this work might be those focused on Kolmogorov complexity (Ryabko and Reznikova, 1996; Hernández-Orallo and Dowe, 2010).

Compression is useful for applications to intelligence in various ways: compression can lead to abstraction, draw attention to underlying structures, help differentiate relevant and

irrelevant portions of its input, and can be a means of increasing processing speed. These are all inherent facets of what we understand of intelligence, and are obviously all interconnected (with the first three points all being methods of achieving increased processing speed).

Furthermore, they are all directly applicable as candidate solutions to some of the weaknesses and limitations of empowerment that were identified in Section 2.7 of Chapter 2 and Section 3.1.1 of Chapter 3:

- All potential future states are considered equal by empowerment.
- It is unclear how to select an appropriate horizon for n-step empowerment.
- Furthermore, it is unclear how to extend the horizon further without hitting the 'boredom' problem or it becoming computationally inviable (let alone biologically plausible).

In Paper 2, I use the information bottleneck method as a form of lossy compression which fits perfectly within the information theoretic framework of empowerment.

4.2 'Outsourcing' to Embodiment and Environment

In Chapter 3 we reviewed how evolution equips an organism with the sensorimotor apparatus to thrive in its niche; specifically, organisms have the perceptual capabilities to detect relevant facets of their environment, and the motor apparatus to manipulate their environment in ways necessary for their success.

However, it may be that this concept actually goes even further, and that the embodiment of an agent or organism can, via its interactions with the environment, not only dictate what actions are available to that agent, but also offset some necessary cognitive processing required for those actions. It seems likely that evolution drives an agent or organism to leverage these structures whenever possible in order to reduce their own metabolic cost and increase fitness.

To illustrate, we can look at the case of a human walking down an incline; a task which mankind has done countless times over the course of its evolution. The physics of the task are quite complex, and yet any healthy adult human can do it without conscious effort. To explain this apparent disparity we could examine Collins et al. (2001), in which a three dimensional walking robot was built without using any electronics; the robot walks down stairs and inclines under the pull of gravity, using only the mechanics in its joints to ascertain the appropriate gait.

A simulated robot, or one using electronics, would need to calculate a variety of non-trivial computations in order to successfully complete such a task. It would need to ascertain the angle of the incline, the angle of each component of its legs and feet, and the relative forces and torque being applied to them, along with a variety of other factors. It is clear to us that humans are not doing any of these calculations, yet they complete the task with ease. The robot in Collins et al. (2001) can also be seen to be completing the task with relative ease.

Essentially the walkbot robot is 'outsourcing' this complex computation to the dynamics and structure of its environment by way of its own embodiment; the robot does not need to do any cognitive processing itself. Usually there is a evolutionary trade-off between the capabilities provided by additional sensorimotor apparatus, and its metabolic cost.

In Pfeifer and Gómez (2009), a similar robot, named *Puppy*, is demonstrated to achieve robust locomotion without the need for sensory feedback from the legs, by the introduction of an artificial 'muscle' (a simple spring). There was a clear distinction between standing and running states, but the controller did not distinguish between them; likewise, it also did not distinguish acceleration or inclination, but a stable gait was observed in this situations. Again, in this instance the 'physiology'/embodiment of *Puppy*, via its interaction with the environment, reduces the computation or processing the agent must do itself, across different phases of action and environmental states.

This concept is sometimes called *embodied cognition* (Varela et al., 1992), and constructs a similar argument whereby the sensorimotor apparatus of an organism is also responsible for

some of the processing. The related idea of *extended cognition* (Clark and Chalmers, 1998) dates back to the 1990s, and proposes a very similar concept whereby the environment itself can play a part of the cognitive process.

We should note that the ability for an embodiment to assist in this way is not only a function of the embodiment, but of that embodiment being in the correct context (i.e. specific part of the correct environment). The walkbot would not help a human kneeling to start a fire, and would be likely less cognitively advantageous if walking down an incline that was submerged in water.

If we follow the hypothesis that sensorimotor adaptations that outsource to the environment reduced metabolic cost, and thus provide an improved trade-off (more 'bang for your buck'), and knowing that evolutionary adaptation would drive biological organisms towards improved adaptedness, we can hypothesise that there is an evolutionary pressure for organisms to 'outsource', where possible, to the environment.

This furthers the hypothesis that an agent or organism's embodiment can indicate not only preferred states, but also preferred actions.

4.3 Paper 2: Introduction

In the conclusion of Paper 1, I discussed the need to better understand how to 'choose an initial strategy when presented with a completely unknown task space' and discussed the selection of a value of n for empowerment.

We can now tie this together with the idea of using compression and the observation that embodiments can encapsulate a degree of cognitive facilities. In Paper 2, I hoped that compression would help empowerment differentiate states from being equal to one another, and indicate states which represented more empowerment than others, and that I could utilise this to help select action sequences. This approach also reduces the processing requirement,

allowing the horizon to be extended further than was previously computationally viable, with the hope that I could address the 'blindness' problem where the horizon was too short to detect salient parts of the environment.

An additional aspect I identified as important when considering which action sequences are more empowered was the presence of noise, specifically any environmental stochasticity that meant transitions between states are not deterministic given an action. This can be an aspect of the environment (e.g wind), part of a game design (e.g. dice rolls), or the presence of other agents (antagonistic or otherwise). When an agent or organism encounters noise they lose a degree of control; such a loss stands in direct contrast to empowerment, so it seems natural than an agent or organism should try to avoid noise in order to maximise empowerment. This aspect felt especially relevant when seeking to impose any sort of compression or bandwidth limitation on empowerment, where a agent must choose between preferred states or action sequences, as there is an obvious pressure to avoid noise.

Paper 2 was published in *Proc. European Conference on Artificial Life 2009* and presented at the *European Conference on Artificial Life* conference (Anthony et al., 2011).

Impoverished Empowerment: 'Meaningful' Action Sequence Generation through Bandwidth Limitation

Tom Anthony, Daniel Polani, Chrystopher L. Nehaniv

Adaptive Systems Research Group, University of Hertfordshire, UK

Abstract. *Empowerment* is a promising concept to begin explaining how some biological organisms may assign *a priori* value expectations to states in taskless scenarios. Standard empowerment samples the full richness of an environment and assumes it can be fully explored. This may be too aggressive an assumption; here we explore impoverished versions achieved by limiting the bandwidth of the empowerment generating action sequences. It turns out that limited richness of actions concentrate on the "most important" ones with the additional benefit that the empowerment horizon can be extended drastically into the future. This indicates a path towards and intrinsic preselection for preferred behaviour sequences and helps to suggest more biologically plausible approaches.

1 Introduction

Methods to provide an agent embodied in an environment with strategies to behave intelligently when given no specific goals or tasks are of great interest in Artificial Life. However, to do this embodied agents require some method by which they can differentiate available actions and states in order to decide on how to proceed. In the absence of no specific tasks or goals it can be difficult to decide what is and is not important to an agent.

One set of approaches examines processing and optimising the Shannon information an agent receives from its environment (Attneave, 1954; Barlow, 1959, 2001; Atick, 1992), following the hypothesis that embodied agents benefit from an adaptive and evolutionary advantage by informationally optimising their sensory and neural configurations for their environment.

Information-based predictions could provide organisms/agents with intrinsic motivation based on *predictive information*(Prokopenko et al., 2006; Bialek et al., 2001; Ay et al., 2008). In this paper we will concentrate on *empowerment* (Klyubin et al., 2005b,a), an information theoretic measure for the external efficiency of a *perception-action loop*.

One shortcoming of empowerment is that whilst it provides behaviours and results which seem to align it with processes that may have resulted from evolution the algorithms used to calculated it tend not to operate using an equally plausible process. It implicitly requires a notion of the richness and full size of the space it searches whatever algorithm is used to determine it. In this paper we thus introduce the assumption of a limit on the richness of the action repertoire.

1.1 Information Theory

First we give a very brief introduction to information theory, introduced by Shannon (1948). The first measure is *entropy*, a measure of uncertainty given by $H(X) = -\sum_x p(x) \log p(x)$ where X is a discrete random variable with values x from a finite set $\mathcal{X}$ and $p(x)$ is the probability that X has the value x. We use base 2 logarithm and measure entropy in *bits*.

If Y is another random variable jointly distributed with X the *conditional entropy* is $H(Y|X) = -\sum_x p(x) \sum_y p(y|x) \log p(y|x)$. This measures the remaining uncertainty about the value of Y if we know the value of X. Finally, this also allows us to measure the *mutual information* between two random variables: $I(X;Y) = H(Y) - H(Y|X)$.

Mutual information can be thought of as the reduction in uncertainty about the variable X or Y, given that we know the value of the other.

1.2 Empowerment

Essentially empowerment measures the channel capacity for the external component of a perception-action loop to identify states that are advantageous for an agent embodied within an environment. It assumes that situations with a high efficiency of the perception-action loop should be favoured by an agent. Based entirely on the sensors and actuators of an agent, empowerment intrinsically encapsulates an evolutionary perspective; namely that evolution has selected which sensors and actuators a successful agent should have, which in turn implies which states are most advantageous for the agent to visit.

Empowerment is based on the information theoretic perception-action loop formalism introduced by Klyubin et al. (2005b,a, 2004), as a way to model embodied agents and their environments. The model views the world as a communication channel; when the agent performs an action, it is injecting Shannon information into the environment, which may or may not be modified, and subsequently the agent re-acquires part of this information from the environment via its sensors.

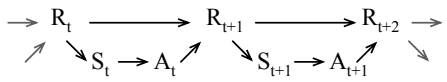


Fig. 1. Bayesian network representation of the perception-action loop.

In Fig.1 we can see the perception-action loop represented by a Bayesian network, where the random variable R_t represents the state of the environment, S_t the state of the sensors, and A_t the actuation selected by the agent at time t. It can be seen that R_{t+1} depends only on the state of the environment at time t, and the action just carried out by the agent.

Empowerment measures the maximum *potential* information flow, this can be modelled by the channel capacity (Shannon, 1948) for a discrete memoryless channel: $C(p(s|a)) = \max_{p(a)} I(A;S)$.

The random variable A represents the distribution of messages being sent over the channel, and S the distribution of received signals. The channel capacity is measured as the maximum mutual information taken over all possible input distributions, $p(a)$, and depends only on $p(s|a)$, which is fixed. One algorithm to find this maximum is the iterative Blahut-Arimoto algorithm (Blahut, 1972).

Empowerment can be intuitively thought of as a measure of how many observable modifications an embodied agent can make to his environment, either immediately, or in the case of n-step empowerment, over a given period of time.

In the case of n-step empowerment, we first construct a compound random variable of the last n actuations, labelled A_t^n. We now need to maximise the mutual information between this variable and the sensor readings at time $t+n$, represented by S_{t+n}. Here we consider empowerment as the channel capacity between these: $\mathfrak{E} = C(p(s_{t+n}|a_t^n)) = \max_{p(a_t^n)} I(A_t^n; S_{t+n})$.

An agent that maximises its empowerment will position itself in the environment in a way as to maximise its options for influencing the environment (Klyubin et al., 2005a).

2 Empowerment with limited action bandwidth

2.1 Goal

We wanted to introduce a bandwidth constraint into empowerment, specifically n-step empowerment where an agent must look ahead at possible outcomes for sequences of actions, and even with a small set of actions these sequences can become very numerous.

An agent's empowerment is bounded by that agent's memory; empowerment measures the agent's ability to exert influence over it's environment and an agent that can perform only 4 distinct actions can have no more than 2 bits of empowerment per step. However, there are two factors which normally prevent empowerment from reaching this bound:

- Noise - A noisy / non-deterministic / stochastic environment means that from a given state an action has a stochastic mapping to the next state. This reduces an agent's control and thus its empowerment.
- Redundancy - Often there are multiple action (or sequences) available which map from a given state to the same resultant state. This is especially true when considering multi-step empowerment: e.g Moving North then West, or moving West then North.

Due to redundancy there are many cases where bandwidth for action sequences can be reduced with little or no impact on achievable information flow. Beyond this there may be scenarios with a favourable trade off between a large reduction in action bandwidth only resulting in a small reduction in empowerment (or utility).

2.2 Scenario

To run tests we constructed a simple scenario; an embodied agent is situated within a 2-dimensional infinite gridworld and has 4 possible actions in any single time step. The actions the agent can execute are North, South, East and West each moving the agent one space into the corresponding cell, provided it is not occupied by a wall. In the scenario the state of the world is solely the position of the agent, which is all that is detected by the agent's sensors.

2.3 Algorithm

The agent to examines all possible sequences for n-step empowerment for small values of n (typically $n < 6$) and then selects a subgroup of the available sequences to be retained.

To do this we use the information bottleneck method (Tishby et al., 1999). Having calculated the empowement we have two distributions: $p(a_t^n)$ is the capacity achieving distribution of action sequences and $p(s_{t+n}|a_t^n)$ is the channel that represents the results of an agent's iteractions with the environment.

We now look for a new "compact" distribution $p(g|a_t^n)$, where g are groups of 'alike' action sequences with $g \in G$ where $|G| \leq |A_t^n|$ and the cardinality of G corresponds to our desired bandwidth limit. A colloquial, though not entirely accurate, way to think of this is as grouping together action sequences that have similar outcomes (or represent similar 'strategies'). The information bottleneck works by first choosing a cardinality for G and then maximising $I(G; S_{t+n})$ (the empowerment of the reduced action set) using S_{t+n} as a relevance variable.

This results in a conditional distribution $p(g|a_t^n)$, from which we must derive a new distribution of our action sequences (with an entropy within the specified bandwidth limit). In order to end up with a subset of our original action sequences to form this new action policy for the agent, we must use an algorithm to 'decompose' the conditional distribution into a new distribution $p(\hat{a}_t^n)$ which has an entropy within the specified bandwidth limit (and usually contains only a subset of the original action sequences).

In the spirit of empowerment, for each g we want to select the action sequences which are most likely to map to that g (i.e the highest value of $p(g|a_t^n)$ for the given g) and provide the most towards our empowerment (i.e the highest value of $I(a_t^n; S_{t+n})$). This results in collapsing strategies to their dominant action sequence and maximises an agent's ability to select between strategies.

2.4 Results

Fig. 2 shows three typical outcomes of this algorithm; in this example we have a bandwidth constraint of 2 bits, operating on sequences of 6 actions. The walls are represented by patterned grey, the starting position of the agent is the light center square, and the selected trajectories by the dark lines with a black cell marking the end location of the sequence. The result that emerges is of interest; the sequences chosen can immediately be seen to be non-trivial and a brief

examination reveals that the end points of each sequence each have only a single sequence (of the available 4,096) that reaches them.

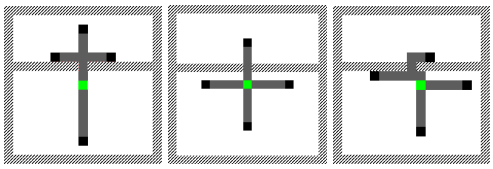


Fig. 2. Typical behaviours where 4 action sequences were selected from 4^6 possibilities.

In section 2.1 we discussed redundancy as one factor which should be eliminated first in order to maintain empowerment whilst reducing bandwidth. If we extrapolate this process of eliminating trajectories to 'easier to reach' states then it follows that, exactly as in Fig. 2, the last states the agent will retain are the entirely unique states that have only a single sequence that reaches them.

It appears that choosing to retain a limited number of explored sequences and this tendency for the agent to value 'unique' sequences indicates a first step towards a solution for extending the sequences beyond what was computationally possible before and may point to a plausible process for a biological organism to undertake. We discuss this in section 3.

2.5 Noise induced behaviour modifications

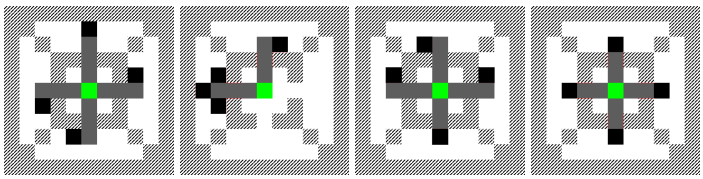


Fig. 3. Randomly selected behaviours; 4 steps with a 2 bit bandwidth constraint. A & B have no noise, C & D have 5% noise per step.

Figures 3 A & B, a 4-step scenario with a bandwidth constraint of 2 bits corresponding to 4 action sequences, show there is not always a neat division of the world into what we would probably recognise as the 4 main 'strategies' (one trajectory into each of the 4 rooms). However, there is no pressure for the agent to do this or to consider the geographical distinctions between states, only for it to select unique end points.

However, with the introduction of noise this changes. Figures 3 C & D show two more randomly selected behaviours from the same scenario but with the introduction of noise, where each action in the sequence has a 5% probability of being replaced with a random action. In order to maintain as much empowerment as possible, the agent must ensure that in attempting one strategy it does not accidentally employ another, and in this environment that translates to being 'blown off course' and adds a drive for a geographical distinction between end states.

Note that some of the sequences appear to be only 3 steps long. This is a strategy employed by the agent, and what is actually happening is the agent uses an action to push against the wall while passing through the doorway, possibly as a way to minimise the effect of noise.

3 Building long action sequences

The current formulation for n-step empowerment utilises an exhaustive search of the action space for $n - steps$. It can be seen that this is a highly unlikely approach for biological organisms to employ, especially for large values of n and in rich environments.

We hypothesise, given that for short sequences of actions it is manageable to cheaply examine all sequences, that we could approach an agent's bandwidth divided into two parts. In section 2.3 we evaluated all possible short sequences in a 'working' memory, then retained only a subset for the agent's 'long term' memory according to our bandwidth constraint.

Following the result above from bandwidth-limited empowerment it became apparent that retaining only a small subset of investigated action sequences lends itself to the idea of then searching further from the final states of such sequences.

This is obvious when applied to the cases where the bandwidth has been constrained just enough to retain empowerment but eliminate all redundancy. It is essentially realising the Markovian nature of such sequence based exploration: when arriving at a state to explore, how you arrived is not of consequence to further exploration. The results, however, seem to suggest that even *beyond* this point of retained empowerment, where the bandwidth severely limits the achievable empowerment and selection of sequences, the iterative approach still produces noteworthy behaviours.

The approach was to set a target length for a sequence, for example 15-step empowerment, then the problem is broken down in to i iterations of n-step empowerment where $n \cdot i = 15$. Standard n-step empowerment is performed, and then the above presented bandwidth-reduction algorithm is run to reduced the action set to a small subset. Each of these action sequences is then extended with n additional steps. These are then again passed through the bandwidth-reduction algorithm and this repeated a total of i times. If we select $n = 5$, $i = 3$ and a bandwidth limit of 4 bits (16 action sequences) then the total sequences evaluated in our gridworld scenario is reduced from 4^{15} to $33,792$, which is a search space more than $3 \cdot 10^4$ times smaller.

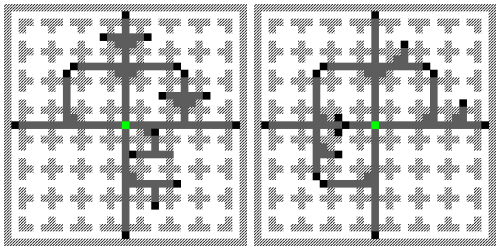


Fig. 4. Iteratively built sequences of 15 steps, with a bandwidth constraint of 4 bits.

Figure 4 shows the results of such a scenario with the selected action sequences and there are several important aspects to note. Firstly, the agent continues to reach certain states that are of obvious consequence, most notably the 4 cardinal directions, but also over half of the furthest reachable corner points. Furthermore the pattern of trajectories has a somehwat 'fractal' nature and appears to divide the search space up systematically. These results are of interest because these states and behaviours are far beyond the horizon of a single iteration of standard n-step empowerment. Space does not permit us to give details but initial results also indicate that interesting locations of the environment, such as door and bridges, are also handled by such iterative sequence building.

4 Discussion

We have identified several challenges to the recently introduced concept of empowerment which endows an agent's environmental niche with a concept distinguishing desirable from less desirable states. Empowerment essentially measures the range in environmental change imprinted by possible action sequences whose number grows exponentially with the length of the sequence. It is virtually impossible to compute it algorithmically for longer sequences, and, likewise, it is implausible that any adaptive or evolutionary natural process would be able to indirectly map this whole range.

Therefore, here we have, consistently with the information-theoretic spirit of our study, applied informational limits on the richness of the action sequences that generate the empowerment. In doing so, we found that: 1. the information bottleneck reduces redundant sequences; 2. in conjunction with the complexity reduction through the collapse of action sequences, particularly "meaningful" action sequences that explore important features of the environment, e.g. principal directions, doors and bridges, are retained, and finally, that *significantly* longer action sequences than before can be feasibly handled. This in itself already suggests insights for understanding the possible emergence of useful long-term behavioural patterns. Note that in this study we have relinquished the computation of empowerment as measure for the desirability of states in favour of filtering out desirable action patterns.

Bibliography

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems*, 3(2):213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.

Ay, N., Bertschinger, N., Der, R., Guettler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *European Physical Journal B*. (Accepted).

Barlow, H. B. (1959). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217–234. The M.I.T. Press.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comp.*, 13(11):2409–2463.

Blahut, R. (1972). Computation of channel capacity and rate distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473.

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

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005a). All else being equal be empowered. In Capcarrère, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life: Proceedings of the 8th European Conference on Artificial Life*, volume 3630 of *Lecture Notes in Artificial Intelligence*, pages 744–753. Springer.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005b). Empowerment: A universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE Press.

Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). Evolving spatiotemporal coordination in a modular robotic system. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., and Parisi, D., editors, *From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy, September 25-29 2006*, volume 4095 of *Lecture Notes in Computer Science*, pages 558–569. Springer.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Tishby, N., Pereira, F., and Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

4.5 Paper 2: Discussion

4.5.1 Iterative Action Sequence Extension

The paper introduced an iterative process for extending the empowerment horizon which, from a practical perspective, is appealing in that it drastically reduces the computation needed for empowerment calculations to more distant horizons. This allowed us to investigate scenarios that were previously not viable due to the computational requirements.

In order to achieve this, I introduced a 'bandwidth constraint' whereby the agent can consider the full richness of options over a few time steps, before retaining only a subset of them as interesting, from which it can then extend the horizon to investigate more deeply. These concepts are similar to the idea of having a 'working memory' and a 'long-term memory'.

From a theoretical perspective it is interesting that when an agent has to think about retaining only a handful of possible initial action sequences from which to then iteratively build on, that a concept of 'sub-goals' can be seen to emerge. Essentially, because the agent has to discard all but a few sequences, it tends to select sequences that are bottleneck points between two separate areas of the state space.

Unfortunately, the algorithm falls short of doing these sub-goals justice, in that due to the absence of any sort of ongoing strategic differentiation (beyond an action sequence being unique, discussed below) action sequences that start off different initially can later converge. This leads to multiple, different, action sequences taking different routes to the same final state.

Whilst identifying different routes to a goal is actually an interesting emerging phenomena, in this instance it would be more appealing (given the limited bandwidth) to instead retain additional sequences that are more strategically differentiated from one another. In the scenarios from the paper, strategic differentiation would be equivalent to the final states being geographically differentiated.

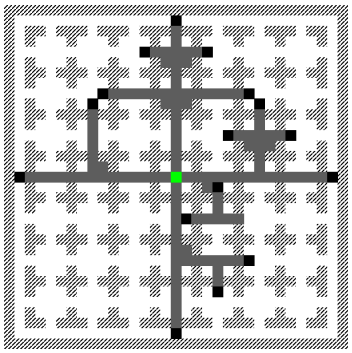


Figure 4.1: *Partial Fig. 4 from Paper 2, showing that an iteratively built sequences of 15 steps, with a bandwidth constraint of 4 bits. Note that there are sets of sequences where the end states are very close to one another, meaning that the collective empowerment of all these states is not as high as it may be were these states more distant from one another.*

When a higher bandwidth was available, meaning less compression and the retention of more available action sequences, then alternative paths to similar states would be appealing.

Looking forward, I established that this iterative approach was promising and potentially powerful, but required an improved method of selecting strategically differentiated states in order to maximise the *collective empowerment* of the states reachable via the retained action sequences.

Collective empowerment is the concept of what the aggregate empowerment would be from multiple positions, measured as a union of the reachable states. It is not actively used, and thus lacks a formal definition, but is a useful conceptual utility to use in analysis and discussion.

Returning to the iterative empowerment approach, in simple deterministic scenarios I wanted to select action sequences that took me to a set of states that, with additional time steps, would allow me to reach a greater set of final states. If two action sequences retained by the agent are geographically nearby then they would share more reachable future with one another, and thus not have such a high collective empowerment. This can be seen in Fig. 4.1.

4.5.2 Redundancy and Unique Sequences

In order to select the most empowering sequences, I applied the information bottleneck method to select a subset of action sequences according to a specified bandwidth, given in bits. The information bottleneck method essentially selects the actions that have the highest mutual information, or the largest 'contribution' towards the channel capacity. This seems an intuitive approach but, as seen in the paper, it leads to an interesting but non-optimal (for our goals) behaviour.

Informally, and in the limited context of this deterministic gridworld scenario, we can imagine each state in the world contributes an equal quantity of empowerment toward the total empowerment value. From there, if we were to assess the empowerment contribution of an action sequence, then each sequence would provide a share of empowerment inversely proportional to the number of other action sequences leading to the same final state.

This leads to some appealing behaviours, such as where the cardinality of retainable action sequences is equal to the number of possible end states, where redundant action sequences are removed and total empowerment maintained. However, when the bandwidth for retaining action sequences is further reduced, such that not all end states can continue to be reachable from a retained action sequence, then the algorithm discards action sequences starting with those with the most number of other action sequences that lead to the same end state. The final action sequences to be retained are therefore those which are entirely unique, in that no other action sequence reaches the same end state.

This has the effect of preferring actions sequences that lead to 'extremes' (geographically or otherwise, depending on the scenario), which inherently has some shortcomings. States that are more 'central' are not represented, which has an effect of narrowing the the strategic breadth of the set of action sequences; this specific effect is offset with iterative extensions where some action sequences then turn back towards these centre states.

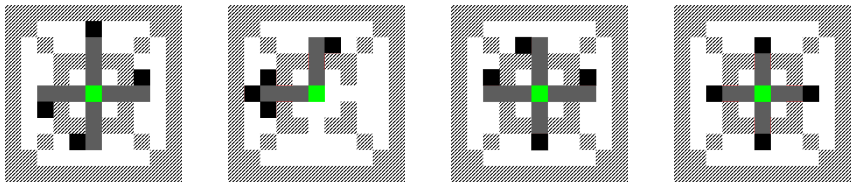


Figure 4.2: *Fig. 3 from the paper, showing a set of selected behaviours, highlighting the effect of noise. The agent was given 4 time steps with a 2 bit bandwidth constraint. A & B have no noise, C & D have 5% noise per step.*

Secondly, the algorithm may choose multiple sequences that lead to unique end states, but these states may not be very different strategically - a point discussed in Section 4.5.1 above.

4.5.3 Noise

The presence of noise has some interesting effects on results of this approach; one might expect that introducing noise into a scenario would adversely effect results. However, if we again look at collective empowerment, then the presence of noise actually improves performance; this comes about because empowerment handles stochasticity 'natively' and proves robust to the presence of noise. In order to prevent that noise from impairing empowerment it selects multiple action sequences that are least likely to end up at the same state, which can be seen in Fig. 4.2, repeated from the paper.

If we look at this in terms of a channel, as per Shannon, then it continues to make sense. In selecting an alphabet, where each signal is a short sequence of primary 'symbols', for communication over a noisy channel, it seems clear that maximally distancing the alphabet sequences in terms of similarity is likely an optimal way to avoid noise turning one valid message into another.

The result of sending one each of the 4 available actions through the 4 available doors seems to represent a Schelling point (Schelling, 1960) - a natural or intuitive solution that humans will converge on.

Furthermore, empowerment demonstrates an additional, unexpected, behaviour in a noisy environment, that aligns with its principle of maintaining as much control as possible. When passing through doorways in a gridworld, the agent would commonly choose to deliberately 'nudge' against a doorway. If such a move went as intended (i.e. noise did not replace the intended action with another) then the agent would remain in the same place, but if the move was affected by the noise then there was an increased chance that the desired state would still be reached.

In the scenario presented in the paper, there is a 5% chance per movement that the intended action is replaced with a randomly selected action from those available. This would sometimes (25% of the time, given the 4 possible actions) be the same action as intended, meaning that the chance of the action being replaced with an alternative is 3.25%. However, when in the doorway then there is a chance the replacement action selected is the opposite cardinal direction (e.g. north instead of south, or east instead of west) which would also push against the door frame and result in the intended state still being reached. This behaviour thereby reduces the effective noise rate from 3.25% to 2.5%.

This result was completely unanticipated, but demonstrates that empowerment is capable of producing non-trivial behaviours, which adapt to circumstances, in a general fashion.

4.5.4 Conclusions and Direction

Essentially, the approach presented in the paper produces self-motivated action selection, but does not do a sufficient job of producing a strategically diverse selection of actions (to maximise future collective empowerment). Furthermore, whilst iterative extension of sequences in a fashion mimicking 'working memory' and 'long-term memory' allows for significantly deeper planning it amplifies the issue of a lack of strategic differentiation.

Those results that included noise demonstrated that empowerment is robust to noise, and also induced an improvement in collective empowerment. This result seeming to be a Schelling

point made it a particularly attractive behaviour, given my motivation. I wanted to find a way to induce this type of behaviour in a more intuitive fashion which wouldn't introduce artificial noise into an environment, which negatively impacts the utility (as empowerment has to also produce behaviours to minimise the effect of that noise).

From here my work focused on improving on the methodology that produced self-motivated action sequence selection, by way of introducing some aspect of strategic planning without compromising the generality or robustness to stochasticity provided by an information-theoretic empowerment based approach.

Chapter 5

Action Selection: Strategies

Action Selection: Strategies

Tactics involve calculations that can tax the human brain, but when you boil them down, they are actually the simplest part of chess and are almost trivial compared to strategy.

— Kasparov (2010)

5.1 Paper 3: Introduction

Paper 2 introduced a method of drastically extending the empowerment horizon, and introduced a form of compression which allowed the empowerment framework to extend to actions.

However, whilst I could evaluate how much of a contribution towards my *current* empowerment an action was providing, I had no method of evaluating how *empowering* an action sequence would be as a measure to whether that action was preferable to another. The ability to measure the contribution of empowerment reduced essentially to uniqueness, which drove interesting behaviours and piqued my interest, but which was not optimal at driving behaviour.

Paper 3 was focused on improving this self-motivated action evaluation to drive viable action sequence policies (in terms of future utility), whilst maintaining the ability to use an iteratively extended horizon model. As part of this, I was motivated to understand the impact that noise was having on the results, as it demonstrated a capacity to help the agent understand that certain future states were 'nearby' one another (in the examples I addressed, this translated neatly to being geographically nearby), but in an uncontrolled fashion.

General Self-Motivation and Strategy Identification: Case Studies based on Sokoban and Pac-Man

Tom Anthony, Daniel Polani, Chrystopher L. Nehaniv

***Abstract*—We use *empowerment*, a recently introduced biologically inspired measure, to allow an AI player to assign utility values to potential future states within a previously un-encountered game without requiring explicit specification of goal states. We further introduce *strategic affinity*, a method of grouping action sequences together to form 'strategies', by examining the overlap in the sets of potential future states following each such action sequence. Secondly, we demonstrate an information-theoretic method of predicting future utility. Combining these methods, we extend empowerment to *soft-horizon empowerment* which enables the player to select a repertoire of action sequences that aim to maintain anticipated utility.**

We show how this method provides a *proto-heuristic* for non-terminal states prior to specifying concrete game goals, and propose it as a principled candidate model for "intuitive" strategy selection, in line with other recent work on "self-motivated agent behaviour". We demonstrate that the technique, despite being generically defined independently of scenario, performs quite well in relatively disparate scenarios, such as a Sokoban-inspired box-pushing scenario and in a Pac-Man-inspired predator game, suggesting novel and principle-based candidate routes towards more general game-playing algorithms.

***Index Terms*—Artificial intelligence (AI), information theory, Games**

I. Introduction

A. Motivation

"Act always so as to increase the number of choices."
- Heinz von Foerster

In many games, including some still largely inaccessible to computer techniques, there exists for many states of that game a subset of actions that can be considered "preferable" by default. Sometimes it is easy to identify these actions, but for many more complex games it can be extremely difficult. While in games such as Chess algorithmic descriptions of the quality of a situation have led to powerful computer strategies, the task of capturing the intuitive concept of the *beauty* of a position, often believed to guide human master players, remains elusive (1). One is unable to provide precise rules for a beauty heuristic, which would need to tally with the ability of master Chess players to appreciate the structural aspects of a game position, and from this identify important states and moves.

Whilst there exist exceedingly successful algorithmic solutions for some games, much of the success derives from a combination of computing power with human explicitly designed heuristics. In this unsatisfactory situation, the core challenge for AI remains: can we produce algorithms able to identify relevant structural patterns in a more general way which would apply to a broader collection of games and puzzles? Can we create an AI player motivated to identify these structures itself?

Game tree search algorithms were proposed to identify good actions or moves for a given state (2; 3; 4). However, it has since been felt that tree search algorithms, with all their practical successes, make a limited contribution in moving us towards 'intelligence' that could be interpreted as plausible from the point of view of human-like cognition; by using brute-force computation the algorithms sidestep the necessity of identifying how 'beauty' and related structural cues would be detected (or constructed) by an artificial agent. John McCarthy predicted this shortcoming could be overcome by brute-force for Chess but not yet Go [1] and criticising that with Chess the solutions were simply 'substituting large amounts of computation for understanding' (5). Recently, games such as Arimaa were created to challenge these shortcomings (6) and provoke research to find alternative methods. Arimaa is a game played with Chess pieces, on a Chess board, and with simple rules, but normally a player with only a few of games experience can beat the best computer AIs.

At a conceptual level, tree search algorithms generally rely on the searching exhaustively to a certain depth. While with various optimizations the search will not, in reality, be an exhaustive search, the approach is unlikely to be mimicking a human approach. Furthermore, at leaf nodes of such a search the state is usually evaluated with heuristics hand-crafted by the AI designer for the specific game or problem.

These approaches do not indicate how higher-level concepts might be extracted from simple rules of the game, or how structured strategies might be identified by a human. For example, given a Chess position a human might consider two strategies at a given moment (e.g. 'attack opponent queen' or 'defend my king') before considering which moves in particular to use to enact the chosen strategy. Tree search approaches do not operate on a level which either presupposes or provides conceptual game structures (the human-made AI heuristics may, of course, incorporate them, but this is then an explicit proviso by the human AI designer).

More recently, Monte Carlo Tree Search (MCTS) algorithms (7) have been developed which overcome a number of the limitations of the more traditional tree search approaches.

T.Anthony, D. Polani and C.L.Nehaniv are with the Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, Hatfield, Herts, AL10 9AB, U.K. e-mail: (research@tomanthony.co.uk,{D.Polani,C.L.Nehaniv}@herts.ac.uk).

[1]Recent progress in Go-playing AI may render McCarthy's pessimistic prediction concerning performance moot, but, the qualitative criticism stands.

MCTS algorithms represent an important breakthrough in themselves, and lead us to a better understanding of tree searching. However, whilst MCTS has significantly extended the potential ability of tree search algorithms, it remains limited by similar conceptual constraints as previous tree search methods.

In the present paper we propose a model that we suggest is more cognitively plausible and yet also provides first steps towards novel methods which could help address the weaknesses of tree search (and may be used in alongside them). The methods we present have arisen from a different line of thought than MCTS and tree search in general. As this is - to our knowledge - the first application of this train of thought to games, at this early stage it is not intended to out-compete state-of-the-art approaches in terms of performance, but rather to develop qualitatively different, alternative approaches which with additional research may help to improve our understanding and approach to game playing AIs.

The model we propose stems from cognitive and biological considerations, and for this purpose we adopt the perspective of intelligence arising from situatedness and embodiment (8) and view the AI player as an agent that is 'embodied' within an environment (9; 10). The agent's actuator options will correspond to the legal moves within the game, and its sensors reflect the state of the game (those parts available to that player according to the relevant rules).

Furthermore, we create an incentive towards structured decisions by imposing a cost on the search/decision process; this is closely related to the concept of *bounded rationality* (11; 12) which deals with decision making when working with limited information, cognitive capacity, and time and is used as a model of human decision-making in economics (13). As natural cost functionals for decision processes, we use information-theoretical quantities; there is a significant body of evidence that such quantities have not only a prominent role in learning theory (14; 15), but also that various aspects of biological cognition can be successfully described and understood by assuming informational processing costs being imposed on organisms (16; 17; 18; 19; 20; 21; 22).

Thus, our adoption of an information-theoretic framework in the context of decisions in games is plausible not only from a learning- and decision-theoretic point of view, but also from the perspective of a biologically oriented high-level view of cognition where pay-offs conferred by a decision must be traded off with the informational effort of achieving them.

Here, more specifically, we combine this "thinking in informational constraints" with *empowerment* (23; 24), another information-theoretic concept generalising the notion of 'mobility' (25) or 'options' available to an agent in its environment. Empowerment can be intuitively thought of as a measure of how many observable changes an embodied agent, starting from its current state, can make to his environment via its subsequent actions. Essentially, it is a measure of mobility that is generalized as it can directly incorporate randomness as well as incomplete information without any changes to the formalism. If noise causes actions to produce less controllable results, this is detected via a lower empowerment value.

This allows one to treat stochastic systems, systems with incomplete information, dynamical systems, games of complete information and other systems in essentially the same coherent way (26).

In game terms, this above technique could be thought of as a type of 'proto-heuristic' that transcends specific game dynamics and works as a default strategy to be applied, before the game-specific mechanics are refined. This could prove useful either independently or as heuristics from genesis which could be used to guide an AI players behaviour in a new game whilst game-specific heuristics were developed during play. In the present paper we do not go as far as exploring the idea of building game-specific heuristics on top of the proto-heuristics, but focus on deploying the method to generate useful behaviour primitives. We demonstrate the operation of proto-heuristics in two game scenarios and show that intuitively 'sensible' behaviours are selected.

B. Information Theory

To develop the method, we require Shannon's theory of information for which we give a very basic introduction. To begin we introduce *entropy*, which is a measure of uncertainty; the entropy of a variable A is defined as:

$$H(A) = -\sum_{a \in A} p(a) \log p(a). \tag{1}$$

where $p(a)$ is the probability that A is in the state a. The logarithm can be taken to any chosen base; in our paper we always use 2, and the entropy is thus measured in *bits*. If S is another random variable jointly distributed with A, the *conditional entropy* is:

$$H(S|A) = -\sum_{a \in A} p(a) \sum_{s \in S} p(s|a) \log p(s|a). \tag{2}$$

This measures the remaining uncertainty about the value of S, if we know the value of A. This also allows us to measure the *mutual information* between two random variables:

$$\begin{aligned} I(A;S) &= H(S) - H(S|A) \\ &= \sum_{a \in A} \sum_{s \in S} p(a,s) \log \left(\frac{p(a,s)}{p(a)\, p(s)} \right) \end{aligned} \tag{3}$$

Mutual information can be thought of as the reduction in uncertainty about one random variable, given that we know the value of the other. In this paper we will also examine the mutual information between a particular value of a random variable with another random variable:

$$I(a;S) = p(a) \sum_{s \in S} p(s|a) \log \left(\frac{p(a,s)}{p(a)\, p(s)} \right). \tag{4}$$

This can be thought of as the 'contribution' by a specific action a to the total mutual information, and will be useful for selecting a subset of A that maximises mutual information.

Finally, we introduce the information-theoretic concept of the *channel capacity* (27). It is defined as:

$$C(p(s|a)) = \max_{p(a)} I(A;S). \tag{5}$$

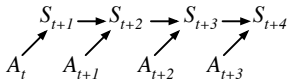


Fig. 1: Bayesian network representation of the perception-action loop.

Channel capacity is measured as the maximum mutual information taken over all possible input (action) distributions, $p(a)$, and depends only on $p(s|a)$, which is fixed for a given starting state s_t. This corresponds to potential maximum amount of information about its prior actions an agent can later observe. One algorithm that can be used to find this maximum is the iterative Blahut-Arimoto algorithm (28).

C. Empowerment

Empowerment, based on the information-theoretic perception-action loop formalism introduced in (24, 29), is a quantity characterizing the "sensorimotor adaptedness" of an agent in its environment. It quantifies the ability of situated agents to influence their environments via their actions.

For the purposes of puzzle-solving and game-play, we translate the setting as follows: consider the player carrying out a move as a sender sending a message, and observing the subsequent state on the board as receiving the response to this message. In terms of Shannon information, when the agent performs an action, it 'injects' information into the environment, and subsequently the agent re-acquires part of this information from the environment via its sensors. Note that in the present paper, we discuss only puzzles and games with perfect information, but the formalism carries over directly to the case of imperfect information game.

For the scenarios relevant in this paper, we will employ a slightly simplified version of the empowerment formalism. The player (agent) is represented by a Bayesian network, shown in Fig. 1, with the random variable S_t the state of the game (as per the player's sensors), and A_t a random variable denoting the action at time t.

As mentioned above, we consider the communication channel formed by the action, state pair A_t, S_{t+1} and compute the channel capacity, i.e. the maximum possible Shannon information that one action A_t can 'inject' or store into the subsequent state S_{t+1}. We define empowerment as this 'motori-sensor' channel capacity:

$$\mathfrak{E} = C(p(s|a)) = \max_{p(a)} I(A; S). \tag{6}$$

If we consider the game as homogenous in time, we can for simplicity ignore the time index, and empowerment only depends on the actual state s_t.

Instead of a single action, it makes often sense to consider an action sequence of length $n > 1$ and its effect on the state. In this case, which we will use throughout most of the paper, we will speak about n-step empowerment. Formally, we first construct a compound random variable of the next n actuations (A_t, A_{t+1}, A_{t+2}, ... , A_{t+n}) = A_t^n. We now maximize the mutual information between this variable and the state at time $t+n$, represented by S_{t+n}. n-step empowerment is the channel capacity between these:

$$\mathfrak{E} = C(p(s_{t+n}|a_t^n)) = \max_{p(a_t^n)} I(A_t^n; S_{t+n}). \tag{7}$$

It should be noted that $\mathfrak{E}$ depends on S_t (the current state of the world) but to keep the notation unburdened, we will always assume conditioning on the current state S_t implicitly and not explicitly write it.

In the present paper we will present two extensions to the empowerment formalism which are of particular relevance for puzzles and games. The first, discussed in section IV, is impoverished empowerment; it sets constraints on the number of action sequences an agent can retain, and was originally introduced in (30). The second, presented in section VI, introduces the concept of a soft horizon for empowerment which allows an agent to use a 'hazy' prediction of the future to inform action selection. Combined these present a model of resource limitation on the actions that can be retained in memory by the player and corresponds to formulating a 'bounded rationality' constraint on empowerment fully inside the framework of information theory.

Prior to the present paper, and (30), empowerment as a measure has solely been used as a utility applied to *states* but in the present paper we introduce the notion of how empowered an action is. In this case *empowered* corresponds to how much a particular action of action sequence contributes towards the empowerment of a state.

Note that we use the Bayesian network formalism in its causal interpretation (31), as the action nodes have a well-defined interventional interpretation — the player can select its action freely. The model is a simpler version of a more generic Bayesian network model of the perception-action loop where the state of the game is not directly accessible, and only partially observable via sensors. The empowerment formalism generalizes naturally to this more general case of partial observability, and can be considered both in the case where the starting state is externally considered ("objective" empowerment landscape) or where it can only be internally observed (i.e. via context, i.e. distinguishing states by observing sensor sequences for which empowerment values will differ see 32; 26). In fact, the empowerment formalism could be applied without change to the more generic *Predictive State Representation* formalism (PSR, see 33)[2].

Here, however, to develop the formalism for self-motivation and strategy identification, we do not burden ourselves with issues of context or state reconstruction, and we therefore concentrate on the case where the state is fully observable. Furthermore, we are not concerned here with with learning the dynamics model itself, but assume that the model is given (which is, in the game case, typically true for one-player games, and for two-player games one can either use a given opponent model or use, again, the empowerment principle to propose a model for the opponent).

[2]Note that this relies on the actions in the entries of the system-dynamics matrix as being interpreted interventionally (i.e. as freely choosable by the agent) in PSR.

Previous results using empowerment in Maze, Box Pushing, and Pole Balancing scenarios demonstrate that empowerment is able to differentiate the preferable (in terms of mobility) states from the less preferable ones (24), correlates strongly with the graph-theoretic measure of closeness centrality (34) in compatible scenarios, and successfully identified the pole being perfectly upright as amongst the most empowered states in various balancing scenarios (26; 35).

II. Related Work

The idea that artifical agents could derive "appropriate" behaviour from their interaction with the environment was implicit already in early work in cybernetics. However, concrete initiatives on how to make that notion precise arose as a consequence of renewed interest in neural controllers, for instance, in the first modern model of artificial curiosity (36). The idea that AI could be applied to generic scenarios with the help of intrinsic motivation models has led to a number of approaches in the last decade. The *autotelic principle* aims at identifying mechanisms which balance skill and challenge as a mechanism for an agent to improve intrinsically (37). Concrete realizations of that are incarnated as *learning progress*, where the progress in acquiring a model of the environment is considered as the quantity to maximize (38); Schmidhuber's *compression progress* framework (39) which bases its measure directly on the progress in compression efficiency in a Kolmogorov-type framework as actions and observations of an agent proceed through time, and which has been extended towards Reinforcement Learning-like frameworks (*AIXI*, 40).

In (41), the model of *intrinsic rewards* emerging from saliency detectors is adopted (which, in turn, may arise in biological agents from evolutionary considerations), and the *infotaxis* exploration model (considered in a biologically relevant scenario) uses (Shannon) information gain about a navigation target as driver for its exploratory behaviour (42).

The notion of predictive information is used in (43) to drive the behaviour of autonomous robots; it is an information-theoretic generalization of the *homeokinesis* principle as to maintain a predictable, but rich (i.e. non-trivial) future behaviour. An overview over a number of principles important for intrinsic motivation behaviours can be found in (44).

Most of the above principles for intrinsic motivation are process-oriented, i.e. they depend both on the environment as well as on the trajectory of the agent through the environment. The latter, in turn, depends on the learning model. In the case of compression progress and *AIXI*, the prior assumptions are relatively minimal, namely a Turing-complete computational model, but true independence from the learning model is only achieved in the asymptotic case.

Infotaxis, as an explorational model, only relies on a model of the environment to induce a locally information-optimal behaviour. Similarly, empowerment does not require any assumptions about learning models; it is not process-oriented, but state-oriented: assume a particular world and agent embodiment, and assume a given empowerment horizon; then, a given state in the world has a well-defined empowerment value, independently of how the agent travels through the world. In particular, empowerment is emphatically not an world exploration model. Though there are some exploration algorithms for model building which are suitable to be plugged into the empowerment computation, the model acquisition phase is conceptually disparate from the empowerment principle at present and we are not aware of a combined treatment of both in a single coherent framework.

In this paper, we generally assume the world and world dynamics to be essentially known. Similar to the intrinsic reward principle by (41), there is the core assumption of an evolutionary "background story" for the relevance of empowerment for a biological organism, but, different from it, empowerment does not assume dedicated saliency detectors, but works on top of the regular perception-action cycle.

III. General Game Playing

In summary, the scenarios reviewed above indicate that empowerment is able to provide a default utility which 1. derives only from the structure of the problem itself and not from an external reward 2. identifies the desirability of states in way that matches intuition and 3. carries over between scenarios of apparently different character.

This makes it a promising candidate to assign a proto-utility to states of a given system, even before a utility (and a goal) have been explicitly specified.

Importantly, empowerment is more than a naive mobility measure; in calculating empowerment for a given state, it incorporates the structure and dynamics of the agent's world and embodiment. In an abstract game scenario, it would be in principle possible to attribute arbitrary labels to actions in different states. However, in biology, there is some evidence that available actions of an organism evolved to match the ecological niche of the organism and simplify its interaction with its environment (8; 45). We propose that a similar match of action set and game dynamics may also be typical for games that humans find attractive to play (similar to the issue of predictability of games (46)); this hypothesis is the basis for us transferring the empowerment formalism from biological models to game-playing.

We believe that empowerment can help move towards a method that could be used for game playing in general [3], there are three primary issues we must first address:

1) There are reasons to suspect that the ability of biological cognition to structure its decision-making process is driven by the necessity to economize its information processing (48). In other words, we postulate that suitable bounded rationality assumptions are necessary to generate structured behaviour. We will represent these assumptions entirely in terms of the language of our information-theoretic framework, in terms of limited 'informational bandwidth' of actions. For games this cognitive cost to processing the environment is especially true where we desire an AI player to play in real-time or at least as fast as a human player.

[3]The problem of "game playing in general" might include, but is not limited to the Stanford AAAI General Game Playing competition (47)

2) For n-step empowerment to be effective in most scenarios, including games, the reliance on a strict horizon depth is problematic and needs to be addressed.
3) The action policy generated by empowerment should identify that different states have different utilities. Naive mobility-like empowerment does not account for the fact that being able to reach some states can be more advantageous that being able to reach others.

In sections IV we will address the first issue. As for issue 2 and 3, it turns out that they are very related to one another; they will be discussed further in sections V and VI.

Finally, in section VII we will bring together all considerations and apply it to a selection of game scenarios.

IV. Impoverished empowerment

In the spirit of bounded rationality outlined above, we modified the n-step empowerment algorithm to introduce a constraint on the bandwidth of action sequences that an agent could retain. We call this modified concept 'impoverished empowerment' (30). This allows us to identify possible favourable trade-offs, where a large reduction in the bandwidth of action sequences has little impact of empowerment.

While in the original empowerment definition, all possible action sequences leading to various states are considered, in impoverished empowerment, one considers only a strongly restricted set of action sequences. Therefore, we need to identify action sequences which are most empowering, i.e those contributing most to the agent's empowerment; how one action sequence can be more empowering than another is a function of the action sequence's stochasticity (does it usually get where it wanted to go), and whether other action sequences lead to the same state (are there other ways to get there).

A. Scenario

To investigate the impoverished empowerment concept we revisited the scenario from (24); a player is situated within a 2-dimensional infinite gridworld and can select one of 4 actions (North, South, East, and West) in any single time step. Each action moves the agent by one space into the corresponding cell, provided it is not occupied by a wall. The state of the world is completely determined by the position of the agent.

B. Impoverished Empowerment Algorithm

This bandwidth reduction works by clustering the available action sequences together into a number of groups, from which a single representative action sequence is then selected. The selected action sequences then form a reduced set of action sequences, for which we can calculate the empowerment.

Stage 1

Compute the empowerment in the conventional way, obtaining a empowerment-maximizing probability distribution $p(a_t^n)$ for all n-step action sequences a (typically with $n < 6$).

Having calculated the empowerment we have two distributions: $p(a_t^n)$ is the capacity achieving distribution of action sequences and $p(s_{t+n}|a_t^n)$ is the channel that represents the results of an agent's interactions with the environment. For conciseness we will write A to represent action *sequences*.

Stage 2

In traditional empowerment computation, $p(a_t^n)$ is retained for all n-step sequences a. Here, however, we assume a bandwidth limitation on how many such action sequences can be retained. Instead of 'remembering' $p(a_t^n)$ for all action sequences a, we *impoverish* $p(a_t^n)$. i.e. we are going to 'thin down' the action sequences to the desired bandwidth limit.

To stay entirely in the information-theoretic framework, we employ the so-called *information bottleneck* method (49; 50). Here, one assumes that the probability $p(s_{t+n}|a_t^n)$ is given, meaning you need a model of what will be possible outcomes for a given action by a player in a given state. In single player games this is easily determined, whereas in multiplayer games we need a model of the other players (we discuss this more in section IX-A).

We start by setting our designed bandwidth limit by selecting a cardinality for a variable G where $|G| \leq |A_t^n|$; we now wish to find a distribution $p(g|a_t^n)$, where g is a group of action sequences with $g \in G$.

The information bottleneck algorithm (see appendix B) can be used to produce this mapping, using the original channel as an input. It acts to minimise $I(G; A_t^n)$ while keeping $I(S_{t+n}; G)$ constant; it can be thought of 'squeezing' the information A_t^n shares with S_{t+n} though the new variable G to maximize the information A_t^n shares with S_{t+n} whilst discarding the irrelevant aspects. By setting a cardinality for G and then running the information bottleneck algorithm we obtain a conditional distribution $p(g|a_t^n)$, which acts as a mapping of actions to groups.

The result of this is action sequences that usually lead to identical states are clustered together to into groups. However, if the number of groups is less than the number of observed states then beyond the identical state action sequences, the grouping is arbitrary, as seen in Fig. 2. This is because there is nothing to imply any relation between states, be it spatial or otherwise - states are only consistently grouped with others that lead to the same state.

Contrary to what might be expected, introducing noise into the environment actually improves the clustering of actions to those that are more 'similar' (in this case spatially). This is due to the possibility to be 'blown off course', meaning the agent sometimes ends up not in the expected outcome state but in a nearby one which results in a slight overlap of outcome states between similar action sequences. However, it is clear that relying on noise for such a result is not ideal and a better solution to this problem is introduced in section VI.

Stage 3

Because our aim is to select a subset of our original action sequences to form the new action policy for the agent, we must use an algorithm to 'decompose' this conditional distribution $p(g|a_t^n)$ into a new distribution of action sequences, which has an entropy within the specified bandwidth limit.

We wish to maximize empowerment, so for each g we select the action sequence which provides the most towards our empowerment (i.e. the highest value of $I(a_t^n; S_{t+n}|g)$). However, when selecting a representative action sequence for a given g we must consider $p(g|a_t^n)$ (i.e. does this action sequence truly represent this group) so we weight on that;

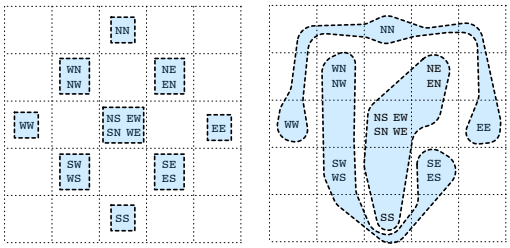


Fig. 2: Visualization of action sequence grouping using Impoverished Empowerment in an empty gridworld with 2-steps; each two character combination (e.g. NW) indicates a 2-step action sequence that leads to that cell from the center cell. Lighter lines represent the grid cells, darker lines the groupings.

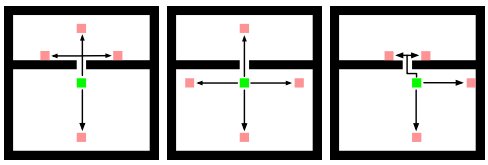


Fig. 3: Typical behaviours where 4 action sequences were selected from 4^6 possibilities. The agent's starting location is shown in green, and its various final locations in pink.

however in most cases the mapping between g and a_t^n is a hard partitioning so this is not normally important. This results in collapsing groups to their 'dominant' action sequence.

C. Impoverishment Results

Figure 3 shows three typical outcomes of this algorithm; in this example we have a bandwidth constraint of 2 bits corresponding to 4 action sequences, operating on sequences with a length of 6 actions; this is a reduction of $4^6 = 4096$ action sequences down to 4. The walls are represented by black, the starting position of the agent is the green center square, and the selected trajectories by the thin arrowed lines with a pink cell marking the end location of the sequence. The result that emerges consistently for different starting states is a set of 'skeleton' action sequences set extending into the state space around the agent. In particular, note that stepping through the doorway which intuitively constitutes an environmental feature of particular salient interest is very often found amongst the 4 action sequences.

Inspection reveals that a characteristic feature of the sequences surviving the impoverishment is the end points of each sequence usually each have a single unique sequence (of the available 4^6) that reaches them.

This can be understood by the following considerations: In order to maintain empowerment whilst reducing bandwidth, the most effective way is to eliminate equivalent actions first since these 'waste' action bandwidth without providing a richer set of end states. States reachable by only one action sequence are therefore advantageous to retain during impoverishment; in Fig. 3, the last action sequences the agent will retain are those leading to states that have only a single unique sequence that reaches them. This is a consequence of selecting the action from each group by $I(a_t^n; S_{t+n})$, and may or may not be desirable; however, in the soft-horizon empowerment model to follow we will see this result disappears.

V. The Horizon

Identifying the correct value for n, for n-step empowerment (i.e. the empowerment horizon depth) is critical for being able to make good use of empowerment in an unknown scenario. A value of n which is too small can mean that states are not correctly differentiated from one another as some options lie beyond the agents horizon. By contrast a value of n which is too large (given the size of the world) can allow the agent to believe all states are equally empowered (30).

Furthermore, it is unlikely that a static value of n would be suitable in many non-trivial scenarios (where different parts of the scenario require different search horizons), and having a 'hard' horizon compounds this.

A. Softening the horizon

We understand intuitively that, when planning ahead in a game, a human player does not employ a hard horizon, but instead probably examines some moves ahead precisely, and beyond that has a somewhat hazy prediction of likely outcomes.

In the soft-horizon empowerment model we use a similar 'softening' of the horizon, and demonstrate how it also helps identify relationships between action sequences which allows the previously presented clustering process to operate more effectively. It allows us to group sets of action sequences together into 'alike' sequences (determined by the overlap in their potential future states), with the resulting groups of action sequences representing 'strategies' . This will be shown later to be useful for making complex puzzles easier to manage for agents. Furthermore, we will show that this horizon softening can help to estimate any ongoing utility we may have in future states, having followed an action sequence. We acknowledge that some future states may be more 'empowered' than others (i.e. lead on to states with more empowerment).

VI. 'Soft-horizon' Empowerment

Soft-horizon empowerment is an extension of the impoverished empowerment model and provides two significant improvements: the clustering of action sequences into groups is enhanced such that the clusters formed represent *strategies*, and it allows an agent to roughly forecast future empowerment following an action sequence. We will show that these features also suggest a solution to having to pre-determine the appropriate horizon value, n, for a given scenario.

A. Split the horizon

Again, we designate a set of actions A, which an agent can select from in any given time step. For convenience we label the number of possible actions in any time step, $|A|$, as c.

We form all possible action sequences of length n, representing all possible action 'trajectories' a player could take in n time steps, such that we have c^n trajectories.

From here, we can imagine a set of c^n possible states the agent arrived in corresponding to the trajectory the agent took, S_{t+n}. It is likely that there are less than c^n unique states, because some trajectories are likely commutative, the game world is Markovian and also because the world is possibly stochastic, but for now we proceed on the assumption that c^n trajectories leads to c^n states (we show how to optimize this in section (30)).

Next, we consider from each of these c^n states what the player could do in an additional m time steps, using the same action set as previously.

We now have for every original trajectory c^n, a set of c^m possible ongoing trajectories. From the combination of these we can create a set of states that represents the outcomes of all trajectories of $n+m$ steps, and label this S_{t+n+m}.

We can form a channel from these states and actions; traditional empowerment's channel would be $p(s_{t+n}|a_t^n)$, corresponding colloquially to 'what is the probability of ending up in a certain state given the player performed a certain action'. With the two trajectories we could form the channel $p(s_{t+n+m}|a_t^{n+m})$, which would be equivalent to if we had simply increased n by m additional steps.

Instead, we create a channel $p(s_{t+n+m}|a_t^n)$, corresponding colloquially to 'what is the probability of ending up in a certain state in $n+m$ steps time if the player performs a given action sequence in the first n steps'. Essentially we are forecasting the potential ongoing future that would follow from starting with a given n-step action sequence.

To do this we need to aggregate and normalise the various distributions of S_{t+n+m} for those which stem from the same original n-step action sequence, a_t^n (their common 'ancestor sequence'). We can calculate this channel:

$$p(s_{t+n+m}|a_t^n) = \frac{\sum_{A_{t+n}^m} p(s_{t+n+m}|a_t^n, a_{t+n}^m)}{|A_{t+n}^m|} \tag{8}$$

where

$$p(s_{t+n+m}|a_t^{n+m}) \equiv p(s_{t+n+m}|a_t^n, a_{t+n}^m) \tag{9}$$

The result of this 'folding back' to ancestor sequences is that the channel now incorporates two important aspects of the initial n-step sequences:

1) each value for a_t^n now has a rough forecast of its future which can be used to approximate a 'future empowerment' value, i.e. what is a players empowerment likely to be after completing the given n-step action sequence, a_t^n.
2) the distribution of potential future states, $S_{t+n+m}|a_t^n$, for different values of a_t^n can be used to compare the potential overlap in the possible futures that follow from those values of a_t^n. This corresponds to how similar they are in terms of strategy, which we call *strategic affinity*.

Point 1 empowers us to differentiate between possible action sequences in terms of utility; naive empowerment is simply counting states whereas this model acknowledges that some potential states will not be as empowered as others. We show how to calculate this forecast of ongoing empowerment in section VI-C.

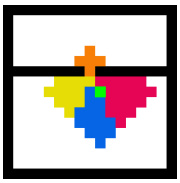


Fig. 4: An example grouping of action sequences, shown here by the colouring of their final states.

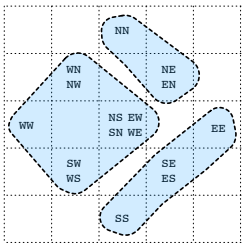


Fig. 5: Visualization of the action sequence grouping from Fig. 2 when the grouping is performed with soft-horizon empowerment. The action sequences now cluster together into 'strategies' formed of similar action sequences that could potentially lead to the same future states.

B. Strategic Affinity

The overlap between the potential futures of each n-step sequence of actions causes them to be grouped together when this channel is fed into the impoverishment algorithm outlined in section IV-B, which brings about the emergence of strategies instead of arbitrary groups previously seen.

The effect of this clustering of action sequences, by their *strategic affinity*, can be illustrated easily in a gridworld as in such an scenario it corresponds closely to geographically close states (see Fig. 5); with more complex worlds such a visualization breaks down but the effect of clustering 'nearby' states remains. An example of such a mapping can be seen in Fig. 4. For many games, this grouping already gives an insight into how tasks may be simplified; either by acting as a coarse representation of the problem or as a tool to identify separate local sub-problems that could be dealt with separately.

Selecting an appropriate bandwidth limit (cardinality) for the number of strategies to be selected is a question not explored in the current paper; we suggest there is rarely a 'correct' answer as selecting a different granularity of strategies will have various different trade-offs.

While soft-horizon empowerment neatly encapsulates strategic affinity, we note that the concept of strategic affinity can be incorporated into other game-playing models, such as MCTS, outside of the empowerment formalism. Combined with a method such as k-means clustering, which does not specify the number of cluster, we hypothesise it would also be possible to use strategic affinity to identify 'natural' strategies.

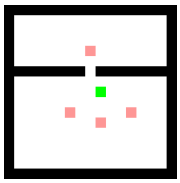


Fig. 6: End states of the 4 action sequences selected to represent the 'strategies' seen in Fig. 4; each is distant from the walls to improve their individual ongoing empowerment.

C. Reducing strategies to actions

We wish to retain only a subset of actions and will use this clustering of action sequences into strategy groups to select a subset of our original action sequences to form the new action policy for the player. To reduce these strategy groups to single action sequences, we select an action sequence from each strategy group that we predict will lead to the most empowered state. We will roughly approximate this forecasted empowerment without actually fully calculating the channel capacity that follows each action sequence.

Earlier we formed the channel $p(s_{t+n+m}|a_t^n, a_{t+n}^m)$ from two separate stages of action sequences, the initial n-steps and the subsequent m-steps. We calculate the channel capacity of this channel which provides a capacity achieving distribution of action sequences $p(a_t^{n+m})$. We now break this channel up into separate channels based on all those where a_t^n is identical, i.e. one channel for each case in which the first n steps are identical. We now have a set of channels, corresponding to each set of m-step sequences that stem from common ancestor sequences. For each of these 'sub-channels' we sum the mutual information for each sequence a_t^{n+m} in this sub-channel with S_{t+n+m}, using the capacity achieving distribution of action sequences calculated above. More formally, $\sum_{a_{t+n}^m \in A_{t+n}^m} I(a_t^n, a_{t+n}^m; S_{t+n+m})$ where $a_t^{n+m} \equiv a_t^n, a_{t+n}^m$. This gives us an approximation of the $(n+m)$-step empowerment for each sequence of n-steps; we can now select, from within each strategy group, those that are most empowered.

As before we must weight this by their likelihood to map to that g (i.e. the highest value of $p(g|a_t^n)$ for the given g), although once again usually these mappings are deterministic so this is unnecessary.

For the mapping shown in Fig. 4 this leads to the selection of action sequences with the end states shown in Fig. 6.

It can be seen that this results in collapsing strategies to the action sequences which are forecast to lead to the most empowered states. Without any explicit goal or reward, we are able to identify action sequences which represent different strategies, and that are forecast to have future utility. The complete soft-horizon empowerment algorithm is presented appendix A.

D. Single-step iterations to approximate full empowerment

One major problem of the traditional empowerment computation is the necessity to calculate the channel capacity in view of a whole set of n-step actions. Their number grows exponentially with n and this computation thus becomes infeasible for large n.

Fig. 7: A Sokoban inspired gridworld scenario. Green represents the player, and purple represents the pushable box, which is blocking the door.

In (30), we therefore introduced a model whereby we iteratively extend the empowerment horizon by 1 step followed by an impoverishment phase that restricts the number of retained action sequences to be equal to the number of observed states. Colloquially this is summed up as 'only remember one action sequence to reach each state'. This is usually sufficient to ensure the player retains full empowerment whilst significantly improving the computational complexity. With this optimisation the computational bounds on empowerment grow with the number of states (usually linear) instead of with the number of action sequences (usually exponential). Space restrictions prevent us from including the algorithm here, but we have used it for the n-phase of empowerment calculations.

E. Alternative Second Horizon Method

An alternative approach, not explored in the present paper, to using the second horizon to predict the future empowerment is to use an equi-distribution of actions over the m-steps forming the second horizon, as opposed to calculating the channel capacity (step in appendix sec:algoappendix). This approximation is algorithmically cheaper, but at the usually at the expense of a less accurate forecast of future empowerment, as well as a less consistent identification of strategies. However, it may prove to be one path to optimising the performance of soft-horizon empowerment.

VII. Game Scenarios and Results

A. 'Sokoban'

Many puzzle games concern themselves with arranging objects in a small space to clear a path, towards a 'good' configuration. Strategy games often are concerned with route finding and similar such algorithms, and heuristics for these often have to be crafted carefully for a particular game's dynamics.

As an example of such games, we examine a simplified box-pushing scenario inspired by Sokoban. In the original incarnation, each level has a variety of boxes which need to be pushed (never pulled) by the player into some designated configuration; when this was completed, the player completes the level and progresses to the next. Sokoban has received some attention for the planning problems it introduces (51; 52), and most pertinent approaches to it are explicitly search-based and tuned towards the particular problem.

We are changing the original Sokoban problem insofar as that in our scenario there are *no* target positions for the boxes, and in fact there is *no* goal or target at all. As stated, we

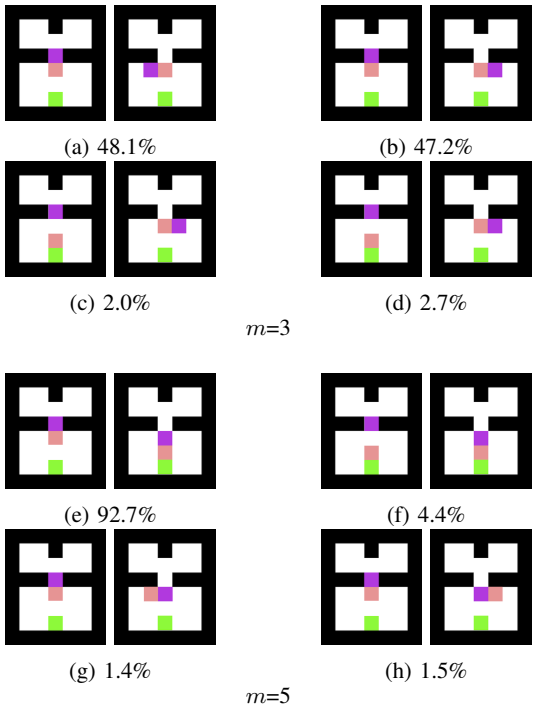


Fig. 8: The distribution of final game states following the selected 4-step action sequences selected by an AI player constrained to 2 action sequences, aggregated from 1000 runs. The pink cells indicate the player's final position.

postulate a critical element for general game playing is self-motivation that precedes the concretization of tasks, thus we adapted the game accordingly.

Figure 7 shows the basic scenario, with a player and a pushable box. Similarly to the earlier gridworlds, the player can move North, South, East and West in any timestep (there is no 'stand still' option). Should the player move into a neighbouring cell occupied by the box, then the box is pushed in the same direction into the next cell; should the destination cell for the box be blocked then neither the player or the box move and the time step passes without any change in the world.

Our intuition would be that most human players, if presented with this scenario and given 4 time steps to perform a sequence of actions on the understanding that an unknown task will follow in the subsequent time steps, would first consider moves that move the box away from blocking the doorway. Humans, we believe, would understand instinctively from observing the setup of the environment, that the box blocks us from the other room and thus moving it gives us more options in terms of places (states) we can reach.

However, it is not obvious how to enable AI players, without explicit goals and no hand-coded knowledge of their environment, to perform such basic tasks as making this identification. Here we approach this task with the fully generic soft-horizon empowerment concept. We use the following parameters: n=4 and m=3, and a bandwidth constraint limiting the player to selecting 2 action sequences from amongst the $4^n = 4^4 = 256$ possible. This constraint was originally selected to see if the AI player would identify both choices for clearing a pathway to the door, and to allow for a possible contrast in strategies.

In Fig. 8 we can see the distribution of final states that were reached. We represent the final states rather than the action sequences that were selected as there are various paths the player can take to reach the same state.

In Fig.8.(a) and Fig.8.(b) we can see the cases that result the majority of the time (95.3%); in one case the box is pushed to the left or right of the door and in the other case the box is pushed into the doorway. The variants in Fig.8.(c) and Fig.8.(d) are almost identical, just the player has moved away from the box one cell.

We can see that two clear strategies emerged; one to clear the path through the door for the player, and a second to push the box through the door (blocking the player from using the doorway). The two options for clearing a path to the doorway (box pushed left or right of the door) are clustered as being part of the same strategy.

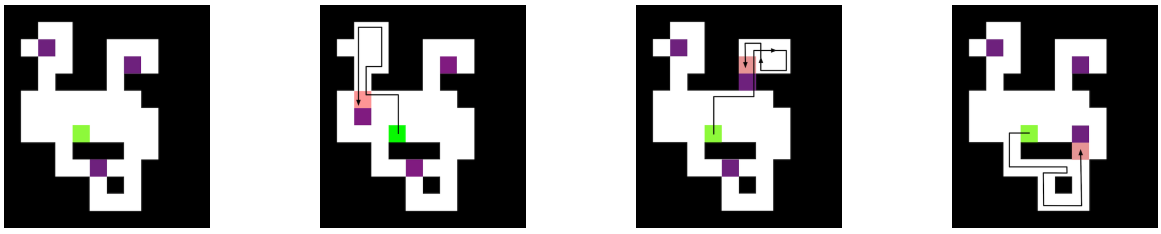
However, it is clear that the types of strategy that can arise, and the ways that action sequences are clustered together, is dependent upon the horizon. If we revisit the same scenario but now adjust the second horizon to be longer, setting $m = 5$ then we can see the results change.

The second row of Fig. 8 show the altered results, and in Fig. 8.(e) we can see that there is a single set of results states that now form the majority of the results (92.7%). We can see that clearing the box to the left or right of the door no longer is part of the strategy; now the player has a horizon of 5 steps it prefers to retain the option of being able to move the box to either the left or the right of the door, rather than committing to one side from the outset. Inspection reveals that occupying the cell below the (untouched) box provides the player with an empowerment of $\mathfrak{E} = log_2 38 = 5.25$ bits rather than $\mathfrak{E} = log_2 31 = 4.95$ bits that would be achieved by clearing the door (in either direction) immediately. The choices that clear the door continue to be clustered together, but the scenario is now 'represented' by the higher empowerment option that emerges with the increased horizon.

Figure 9 shows a more complex scenario, with multiple boxes in the world, all 'trapped' in a simple puzzle. Again, there is no explicit goal; for each of the 3 boxes there exists a single unique trajectory that will recover the box from the puzzle without leaving it permanently trapped. Note that trapping any box immediately reduces the player's ability to control its environment and costs it some degrees of freedom (in the state space) afforded by being able to move the box.

The scenario is designed to present multiple intuitive 'goals', which are attainable only via a very sparse set of action sequences. With a horizon of 14 steps there are 268 million possible action sequences (leading to 229 states), of which 13 full retrieve a box. Note that box 2 (top right) cannot be fully retrieved (pass through the doorway) within 14-steps, and that box 1 (top left) is the only box that could be returned to its starting position by the player.

Table I shows the results when allowing the AI player to select 4 action sequences of 14-steps with an extended horizon



(a) Scenario (b) Box 1 (c) Box 2 (d) Box 3

Fig. 9: A second Sokoban scenario, with multiple boxes in purple and the player's starting position in green. Solutions shown represent the fullest possible recovery for the 3 boxes, along with example action-sequences for recovery. Being 1-step short of these solutions is considered partial recovery (not shown).

	Box 1	Box 2	Box 3
Full	81%	43%	69%
Partial/Full	93%	53%	82%

TABLE I: Percentage of runs in which each of the boxes was recovered. We can see the importance of a long enough horizon; box 2 (which cannot be retrieved completely from the room) is recovered less often than the other boxes.

of m=5 steps, averaged over 100 runs. An explicit count of the different paths to the doorway for each box's puzzle room reveals that there are only 6 action sequences that fully retrieve the box to the main room for each of the top 2 boxes, and only one still for the bottom box.

The results indicate the ability of soft-horizon empowerment to discover actions that lead to improved future empowerment. Furthermore, in every run all 3 boxes were moved in some way, with 35% of cases resulting in all 3 boxes are retrieved, and in 58% at least two boxes are retrieved, leading to indications that the division of the action-sequences into strategies is a helpful mechanism towards intuitive goal identification.

B. Pac-Man-inspired Predator-Prey Maze Scenario

Pac-Man and its variants have been studied previously, included using a tree search approach (53), but the aim of the current paper is not to attempt to achieve the performance of these methods but rather to demonstrate that, notwithstanding their genericness, self-motivation concepts such as soft-horizon empowerment are capable of identifying sensible goals of operation in these scenarios on their own and use these to perform the scenario tasks to a good level.

Thus, the final scenario we present is a simplified predator-prey game based on a simplified Pac-Man model; rather than having pills to collect, and any score, the game is simplified to having a set of ghosts that hunt the player and kill him should they catch him. The 'score', which we will measure, is given simply by the time-steps that the player survives before he is killed; however it is important to note that our algorithm will not be given an explicit goal, rather the implicit aim is simply survival. If humans play the game for a few times, it is plausible to assume (and we would also claim some anecdotal evidence for that) that they will quickly decide that survival is the goal of the game without being told. Choosing survival as your strategy is a perfectly natural decision; assuming no further knowledge/constraints beyond the game dynamics, and a single-player game, anything that may or may not happen later has your survival in the present as its precondition.

In the original Pac-Man game, each ghost uses a unique strategy (to add variation and improve the gameplay) and they were not designed to be ruthlessly efficient; the ghosts in our scenario are far more efficient and all use the same algorithm. Here, in each timestep, the player makes a move (there is no 'do nothing' action, but he can indirectly achieve it by moving towards a neighbouring wall), and then the ghosts, in turn, calculate the shortest path to his new location and move. Should multiple routes have the same distance, then the ghosts randomly decide between them. They penalise a route which has another ghost already on it by adding d extra steps to that route; setting $d = 0$ results in the ghosts dumbly following one another in a chain which is easy for the player. Increasing the value makes the ghosts swarm the player more efficiently. For the present results we use $d = 8$ which is a good compromise between ghost efficiency and giving the player sufficient chance to survive long enough to allow different values for n and m to differentiate.

The maze setup we used is shown in Fig. 10, and the location of the 3 ghosts can be seen. Having only 3 ghosts is another compromise for the same reasons as above; using 4 ghosts usually resulted in the player surviving not long enough to get meaningful variance in the results generated with different parameter sets.

The player has a model of the ghosts' algorithm and thus can predict their paths with some accuracy, and is being allowed 4 samples of their possible future positions (which are stochastic given the possibility that for one or more ghosts the path lengths coincide) for a given move of his. However, once no equal routes are present then 1 sample is perfect information, but once one or more ghosts has one or more equal length paths, then the sampling becomes less accurate and may lose information about the possible future ghost moves.

The game begins with the player's first move, and continues until he is caught by any of the ghosts; at this point the player is 'dead' and is no longer allowed to move. However, there is

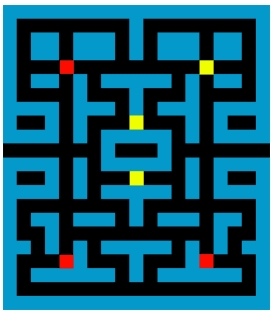


Fig. 10: Pac-Man-inspired scenario, showing the three possible starting positions of the player (yellow) in the center, and of the starting positions of each of the 3 ghosts (red).

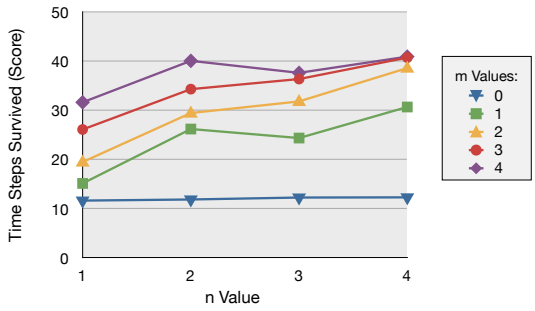


Fig. 11: The player's 'score' for different parameter sets, averaged over 3 different starting positions for the player, and 100 games per data point. It highlights that with no second horizon (m=0) performance does not improve as the first horizon (n) increases.

no special 'death' state in our model; once caught, the player is no longer allowed to move but can still observe the game state (which quickly stops changing anyway, as the ghosts, having swarmed the player, no longer move).

Using the above algorithm, with a cardinality of strategies set to 1 to pick a single action sequence, we observe that the player flees from the ghosts once they come into his horizon; this result from the fact that his future control over the state of the game would drop to zero should he be caught. Death translates directly into the empowerment concept by a vanishing empowerment level. Figure 11 shows the results for various parameter sets, for 3 different starting positions; for each combination of starting position, n value and m value we ran 100 games, then averaged the number of time steps survived over the starting positions for a final average 'score' for each combination of n and m.

Firstly, it can be seen that for $m = 0$, which is equivalent to 'standard' empowerment (23; 24) and does not make use of any features of the algorithm presented that increasing the value of n has no impact on the player's performance. Without a second horizon and thus some measure of his control over the game in the future (beyond the first horizon) there is no pressure to maintain that control. Colloquially, we could say the player only cares about his empowerment in the present moment, not at all about the future. Being able to reach a future state in which he is dead or trapped seems just as good as being able to reach a future state in which he still has a high empowerment; the result is he does not even try to avoid the ghosts and is easily caught.

Once the player has even a small ongoing horizon with $m = 1$ it is easy to see the increase in performance, and with each increase in m performance improves further as the player is better able to predict his future state beyond what action sequence he plans to perform next. For all cases where $m > 0$ it can be seen there is a general trend that increasing n is matched with increasing performance, which would be expected; planning further ahead improves your chances to avoiding the ghosts and finding areas of continued high empowerment.

Note that $n = 2$ performs well and outside of the fit of the other results; this seems, from inspection of individual runs, to be an artefact of the design of the world and the properties of one of the three starting positions, and does not persist that strongly when the starting position is changed. This highlights how a given structure or precondition in a world, which is not immediately observable, could be exploited by specific, hand-crafted AI approaches unique to that exact situation but would be difficult to transfer to other scenarios. The results are shown again, separately for each m value in Fig. 12.

One interesting non-trivial behaviour that consistently emerged from the soft-horizon empowerment algorithm in this scenario was a kiting technique the player would use to 'pull ghosts in'; his employed strategy favoured having a ghost in the immediate cell behind him (this makes that particular ghosts behaviour completely predictable and not only reduces the players's uncertainty about the future but also increases his ability to control it - this includes having a persistent option to commit suicide in a controlled manner at any point). Therefore, the player could often be observed moving back and forth between two cells waiting for a nearby ghost to get to such a position; however, in our observations this did not happen when other ghosts are nearby which would result in the danger of the player being surrounded. This behaviour is not something that would seem intuitive to a human player in this scenario (but humans employ kiting as a method in other games), and whether skirting danger in such a way is desirable in other scenarios is hard to predict.

VIII. Comparison to Mobility

In order to highlight some important differences between soft-horizon empowerment and a greedy mobility algorithm of similar complexity, we present a brief example from the gridworld scenario seen earlier. We created a simple algorithm that samples the world in the same way, and operates with the same goal as soft-horizon empowerment: to select a specified number of actions to maximise utility. The algorithm works thus:

1) Sample the results of performing all possible n-step actions sequences to produce $p(s_{t+n}|a_t^n)$
2) From all reachable states (S_{t+n}), calculate the average mobility (denoted $U(S_{t+n})$) (by sampling) from that state achievable in m-steps

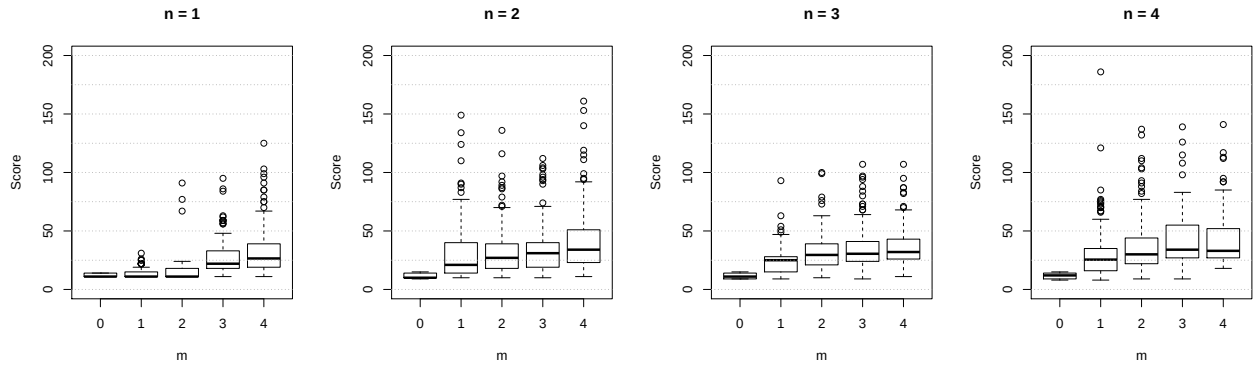


Fig. 12: Boxplot showing quartiles,medians, maximum and minimums for scores across 100 games per pair of horizon values (n,m) for the starting position indicated in Fig 10. The whiskers show 1.5x the interquartile range. It can be observed that a combined horizon of 3-4 is necessary to survive, and then there is a trend towards improved scores as m increases, but as n becomes larger this trend plateaus as performance no longer significantly improves.

Fig. 13: Gridworld scenario. The green cell represents the starting position of the player; the player starts atop a wall.

3) For each value of a_t^n, calculate the average mobility the player would achieve, $p(s_{t+n}|a_t^n) \cdot U(S_{t+n})$
4) Select, according to a specified cardinality, those actions with the highest expected mobility (where there are multiple actions with equally maximal expected mobility, pick a subset randomly)

This results in a set of n-step actions sequences that have an expected mobility over $n+m$-steps in a similar fashion to soft-horizon empowerment, and like empowerment can operate in stochastic scenarios.

A. *Gridworld scenario*

This gridworld scenario we present here, shown in Fig. 13, operates identically to the gridworld scenarios earlier, but in this instance there is no box for the agent to manipulate. The player starts on a wall along which it cannot move; the player can move east or west and 'fall' into the room on either side, but it cannot then return to the wall. Whilst on the wall the player can effectively stay still by trying to move north or south.

We ran both soft-horizon empowerment and the greedy mobility algorithm with $n = 2$, $m = 2$ and an target action sequence cardinality of 1, such that the output of each algorithm would be a 2-step action sequence selected to maximise empowerment or mobility. For each method, 1000 runs were performed and the results are shown in table II.

All of the actions selected by the greedy mobility algorithm have an expected mobility of 9 moves, and all those moves also lead to a state where $\mathfrak{E} = 3.17$ bits. However, due to the way in which the soft-horizon empowerment algorithm forecasts future empowerment (in the second horizon), it favours moving away from the wall.

Action	Empowerment	Greedy Mobility
EE	51.8%	7.4%
WW	48.2%	7.7%
NN/SS/NS/SN	-	34.2%
EN/ES	-	25.0%
WN/WS	-	25.7%

TABLE II: Distribution of action sequences selected by each method (each over 1000 runs). Action sequences leading to the same state have been grouped. No noise.

Action	Empowerment	Greedy Mobility
EE	-	7.4%
WW	100.0%	8.8%
NN/SS/NS/SN	-	35.5%
EN/ES	-	18.1%
WN/WS	-	30.2%

TABLE III: Distribution of action sequences selected by each method (each over 1000 runs), with noise in the eastern room.

We now introduced some noise into the environment; making it so in the eastern room there was a 50% chance that, for any run, all directions are rotated (N→S, S→E, E→W, W→N). Again, 1000 runs were performed and the results are shown in table III.

The change can be seen clearly; empowerment immediately

adapts and switches to always favouring the western room where it has more control, whereas greedy mobility does not significantly change the distribution of actions it selects. This behaviour is a critical advantage of empowerment over traditional mobility; the rotation of actions does nothing to decrease mobility as each action within a specific run is deterministic. When beginning a new run it is impossible to predict whether the actions would be inverted or not, so whilst there is no decrease in mobility, there is a decrease in *predictability* which negatively impacts empowerment.

Whilst empowerment can be intuitively thought of as a stochastic generalization of mobility, it is actually not exactly the case in many instances; it is possible to encounter stochasticity with no reduction in mobility, but stochasticity is reflected in empowerment due to its reducing of a players control (over their own future).

IX. Discussion

The presented soft-horizon empowerment method exhibits two powerful features, both of which require no hand-coded heuristic:

- the ability to assign sensible 'anticipated utility' values to states where no task or goal has been explicitly specified.
- accounting for the strategic affinity between potential action sequences, as implicitly measured by the overlap in the distribution of their potential future states (naively this can be thought of as how many states that are reachable from A are also reachable from B within the same horizon). This allows a player to select a set of action sequences that fit within a specific bandwidth limit whilst ensuring that they represent a diversity of strategies.

This clustering of action-sequences allows strategic problems to be approached with a coarser grain; by grouping sets of actions together into a common strategy, different strategies can be explored without requiring that every possible action sequence is explored. We believe that such a grouping moves towards a 'cognitively' more plausible perspective which groups strategies a priori according to classes of strategic relevance rather than blindly evaluating an extremely large number of possible moves. Furthermore, by modifying the bandwidth, the concept of having strategies with differing granularities (i.e. 'attack' versus 'attack from the north' and 'attack from the south' etc.) emerges; it has previously been shown that there is a strong urge to compress environments and tasks in such a way (54).

Before, however, we go into a more detailed discussion of the approach in the context of games, some comments are required as to why a heuristic which is based only on the structure of a game and does not take the ultimate game goal into account, can work at all. This is not obvious and seems, on first sight, to contradict the rich body of work on reward-based action selection (grounded in utility theory/reinforcement learning etc.).

To resolve this apparent paradox, one should note that for many games, the structure of the game rules already implicitly encodes partial aspects of the ultimate tasks to a significant degree (similarly to other tasks (55)). For instance, Pac-Man by its very nature is a survival game. Empowerment immediately reflects survival, as a 'dead' player loses all empowerment.

A. Application to Games

In the context of tree search, the ability to cluster action-sequences into strategies introduces the opportunity to imbue game states and corresponding actions with a relatedness which derives from the intrinsic structure of the game and is not externally imposed by human analysis and introspection of the game.

The game tree could now be looked at from a higher level, where the branches represent strategies, and the nodes represent groups of similar states. It is possible to foresee pruning a tree at the level of thus determined strategies rather than individual actions, incurring massive efficiency gains. More importantly, however, these strategies emerge purely from the structure of the game rather than from an externally imposed or assumed semantics.

In many games, it is reasonable to assume having perfect knowledge of transitions in the game state given a move. However, note that the above model is fully robust to the introduction of probabilistic transitions, be it through noise, incomplete information or simultaneous selection of moves by the opponent. The only precondition is is the assumption that one can build a probabilistic model of the dynamics of the system. Such opponent or environment models can be learned adaptively (35; 56). The quality of the model will determine the quality of the generated dynamics, however, we do not investigate this here further.

We illustrated the efficacy of this approach using two scenarios. Importantly, the algorithm was not specifically crafted to suit the particular scenario, but is generic and transfers directly to other examples.

In the Pac-Man-inspired scenario, we demonstrated that acting to maintain anticipated future empowerment is sufficient to provide a strong generic strategy for the player. More precisely, the player, without being set an explicit goal, made the 'natural' decision to flee the ghosts. This behaviour derives from the fact that empowerment is by its very nature a 'survival-type' measure, with death being a 'zero-empowerment' state. With the second horizon's forecast of the future, the player was able to use the essential basic empowerment principle to successfully evade capture for extended periods of time.

We presented several Sokoban-inspired scenarios; the first, smaller, scenario presented a doorway that was blocked by a box, with human intuition identifying clearing the doorway as a sensible idea. We saw that soft-horizon empowerment identified clearing the doorway as a good approach for maximising future utility, and also selected an alternative strategy of pushing the box through the doorway. It was interesting to see that soft-horizon empowerment identified clearing the door with the box to the left or to the right as part of the same strategy, as opposed to pushing the box through the door. This scenario also highlighted how the algorithm differently differentiates strategies based on its horizon limitations.

The second scenario presented 3 trapped boxes each requiring a 14-step action sequence to 'retrieve' from the trap. A

human introspecting the problem could deduce the desirable target states for each box (freeing them so they could be moved into the main room). With a total of 268×10^6 possible action sequences to choose from, and lacking the *a priori* knowledge determining which states should be target states, the algorithm reliably selects a set of action sequences which includes an action sequence for retrieving each of the boxes. Not only are the target states identified as being important but the possible action sequences to recover each of the different boxes are identified as belonging to a different strategy.

The importance of this result lies in the fact that, while again, the approach used is fully generic, it nevertheless gives rise to distinct strategies which would be preferred also based on human inspection. This result is also important for the practical relevance of the approach. The above relevant solutions are found consistently, notwithstanding the quite considerable number and depth of possible action sequences. We suggest that this may shed additional light on how to construct cognitively plausible mechanisms which would allow AI agents to preselect candidates for viable medium-term strategies without requiring full exploration of the space.

The final Sokoban example introduced noise into the environment and compared empowerment to a simple mobility algorithm. It highlighted a distinct advantage of empowerment over mobility in that empowerment identifies a reduction in control and is able to respond appropriately.

B. Relation to Monte-Carlo Tree Search

The presented formalism could be thought of, in certain circumstances as just dividing a transition table into two halves and using forecasts of probabilities of encountering states in the second half to differentiate those states in the first half and assign them some estimated utility. The information-theoretic approach allows this to be quantifiable and easily accessible to analysis. However, we believe the technique presented would work using other methodologies and could be combined with other techniques in the medium-term. One important example of where it could be applied would be in conjunction with a Monte-Carlo Tree Search approach, and we would like to discuss below how the formalism presented in this paper may provide pathways to address some weaknesses with MCTS.

MCTS has been seen to struggle with being a global search in problems with a lot of 'local structure' (57). An example for this is a weakness seen in the Go progam Fuego, which is identified as having territorially weak play (58) because of this problem. Some method which clusters of action sequences into strategies, where the stategic affinity 'distance' between the subsequent states is low, might allow for the tree search to partially operate at the level the strategies instead of single actions and this could help in addressing the problem.

The second aspect of MCTS which has led to some criticism is that it relies on evaluating states to the depth of the terminal nodes they lead to in order to evaluate a state. It is possible that the 'folding back' model of empowerment presented in this paper could be used as a method to evaluate states in an MCTS, which may operate within a defined horizon when no terminal states appear within that horizon. In this way the search could be done without this terminal state requirement, and this might allow a better balance between the depth of the search versus its sparsity. This, of course, would make use of the implicit assumption of structural predictability underlying our formalism.

C. Computational Complexity

We are interested in demonstrating the abilities of the soft-horizon method, but in the present paper we did not aim yet for an optimization of the algorithm. As such, the unrefined soft-horizon algorithm is still very time-consuming.

Previously the time complexity of empowerment was exponential with respect to the horizon, n, until the impoverish-and-iterate approach was introduced in (30) (which is linear with the respect to the number of states encountered).

The present paper introduces soft-horizon empowerment and a second horizon m, and currently the time complexity of the soft-horizon algorithm is exponential with respect to m. We have not yet attempted to apply the iterative impoverishment approach to soft-horizon empowerment, but we expect that this or a similar approach would provide significant improvements.

In continuous scenarios, where early empowerment studies used to be extremely time-consuming, recently developed approximation methods for empowerment allowed to reduce computation time by several orders of magnitude (59).

X. Conclusion

We have proposed soft-horizon empowerment as a candidate for solving implicit 'problems' which are defined by the environment's dynamics without imposing an externally defined reward. We argued that these cases are of a type that are intuitively noticed by human players when first exploring a new game, but which computers struggle to identify.

Importantly, it is seen that this clustering of action sequences into strategies determined by their strategic affinity, combined with aiming to maintain a high level of empowerment (or naive mobility in simpler scenarios) brings about a form of 'self-motivation'. It seems that setting out to maximize the agent's future control over a scenario produces action policies which are intuitively preferred by humans. In addition, the grouping of action sequences into strategies ensures that the 'solutions' produced are diverse in nature, offering a wider selection of options instead of all converging to micro-solutions in the same part of the problem at the expense of other parts. The philosophy of the approach is akin to best preparing the scenario for the agent to maximize its influence so as to react most effectively to an as yet to emerge goal.

In the context of general game-playing, to create an AI that can play new or previously un-encountered games, it is critical to shed its reliance on externally created heuristics (e.g. by humans) and enable it to discover its own. In order to do this, we propose that it will need a level of self-motivation and a general method for assigning preference to states as well as for identifying which actions should be grouped into similar strategies. Soft-horizon empowerment provides a starting point into how we may begin going about this.

REFERENCES

[1] S. Margulies, "Principles of Beauty," *Psychological Reports*, vol. 41, pp. 3–11, 1977.

[2] J. von Neumann, "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, vol. 100, no. 1, pp. 295–320, 1928.

[3] C. E. Shannon, "Programming a computer for playing chess," *Philosophical Magazine*, vol. 41, no. 314, p. 256–275, 1950.

[4] A. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal*, vol. 11, pp. 601–617, 1967.

[5] J. McCarthy, "What is Artificial Intelligence?" 2007. [Online]. Available: http://www-formal.stanford.edu/jmc/whatisai/

[6] O. Syed and A. Syed, "Arimaa - a new game designed to be difficult for computers," *International Computer Games Association Journal*, vol. 26, pp. 138–139, 2003.

[7] G. Chaslot, S. Bakkes, I. Szita, and P. Spronck, "Monte-Carlo Tree Search: A New Framework for Game AI," in *AIIDE*, 2008.

[8] R. Pfeifer and J. C. Bongard, *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books)*. The MIT Press, 2006.

[9] F. J. Varela, E. T. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, new edition ed. The MIT Press, Nov. 1992.

[10] T. Quick, K. Dautenhahn, C. L. Nehaniv, and G. Roberts, "On Bots and Bacteria: Ontology Independent Embodiment," in *Proc. of 5th European Conference on Artificial Life (ECAL)*, 1999, pp. 339–343.

[11] H. A. Simon, *Models of man: social and rational; mathematical essays on rational human behavior in a social setting*. New York: Wiley, 1957.

[12] O. E. Williamson, "The Economics of Organization: The Transaction Cost Approach," *The American Journal of Sociology*, vol. 87, no. 3, pp. 548–577, 1981.

[13] C. A. Tisdell, *Bounded rationality and economic evolution : a contribution to decision making, economics, and management*. Edward Elgar, Cheltenham, UK, 1996.

[14] D. A. McAllester, "PAC-Bayesian Model Averaging," in *In Proceedings of the Twelfth Annual Conference on Computational Learning Theory*. ACM Press, 1999, pp. 164–170.

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

[16] F. Attneave, "Some Informational Aspects of Visual Perception," *Psychological Review*, vol. 61, no. 3, pp. 183–193, 1954.

[17] H. B. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," in *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, W. A. Rosenblith, Ed. The M.I.T. Press, 1959, pp. 217–234.

[18] ——, "Redundancy Reduction Revisited," *Network: Computation in Neural Systems*, vol. 12, no. 3, pp. 241–253, 2001.

[19] J. J. Atick, "Could Information Theory Provide an Ecological Theory of Sensory Processing," *Network: Computation in Neural Systems*, vol. 3, no. 2, pp. 213–251, May 1992.

[20] M. Prokopenko, V. Gerasimov, and I. Tanev, "Evolving Spatiotemporal Coordination in a Modular Robotic System," in *From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy, September 25-29 2006*, S. Nolfi, G. Baldassarre, R. Calabretta, J. Hallam, D. Marocco, J.-A. Meyer, and D. Parisi, Eds., vol. 4095. Springer, 2006, pp. 558–569.

[21] W. Bialek, I. Nemenman, and N. Tishby, "Predictability, Complexity, and Learning," *Neural Comp.*, vol. 13, no. 11, pp. 2409–2463, 2001.

[22] N. Ay, N. Bertschinger, R. Der, F. Guettler, and E. Olbrich, "Predictive Information and Explorative Behavior of Autonomous Robots," *European Physical Journal B*, 2008.

[23] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Empowerment: A Universal Agent-Centric Measure of Control," in *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, vol. 1. IEEE Press, 2005, pp. 128–135.

[24] ——, "All Else Being Equal Be Empowered," in *Advances in Artificial Life: Proceedings of the 8th European Conference on Artificial Life*, ser. Lecture Notes in Artificial Intelligence, M. S. Capcarrère, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, Eds., vol. 3630. Springer, Sep 2005, pp. 744–753.

[25] E. Slater, "Statistics for the chess computer and the factor of mobility," *Information Theory, IRE Professional Group on*, vol. 1, no. 1, pp. 150 –152, Feb 1953.

[26] 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, 12 2008.

[27] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[28] R. Blahut, "Computation of Channel Capacity and Rate Distortion Functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, Jul 1972.

[29] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Organization of the Information Flow in the Perception-Action Loop of Evolved Agents," in *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, R. S. Zebulum, D. Gwaltney, G. Hornby, D. Keymeulen, J. Lohn, and A. Stoica, Eds. IEEE Computer Society, 2004, pp. 177–180.

[30] T. Anthony, D. Polani, and C. L. Nehaniv, "Impoverished Empowerment: 'Meaningful' Action Sequence Generation through Bandwidth Limitation," in *Proc. European Conference on Artificial Life 2009*. Springer, 2009.

[31] J. Pearl, *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press, 2000.

[32] P. Capdepuy, D. Polani, and C. L. Nehaniv, "Constructing

the Basic Umwelt of Artificial Agents: An Information-Theoretic Approach," 2007, pp. 375–383.

[33] S. Singh, M. R. James, and M. R. Rudary, "Predictive State Representations: A New Theory for Modeling Dynamical Systems," in *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI)*, 2004, pp. 512–519.

[34] T. Anthony, D. Polani, and C. L. Nehaniv, "On Preferred States of Agents: how Global Structure is reflected in Local Structure," in *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, S. Bullock, J. Noble, R. Watson, and M. A. Bedau, Eds. MIT Press, Cambridge, MA, 2008, pp. 25–32.

[35] T. Jung, D. Polani, and P. Stone, "Empowerment for continuous agent-environment systems," *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, vol. 19, pp. 16–39, February 2011.

[36] J. Schmidhuber, "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, J. A. Meyer and S. W. Wilson, Eds. MIT Press/Bradford Books, 1991, pp. 222–227.

[37] L. Steels, "The Autotelic Principle," in *Embodied Artificial Intelligence: Dagstuhl Castle, Germany, July 7-11, 2003*, ser. Lecture Notes in AI, F. Iida, R. Pfeifer, L. Steels, and Y. Kuniyoshi, Eds. Berlin: Springer Verlag, 2004, vol. 3139, pp. 231–242.

[38] P. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, pp. 265–286, 2007.

[39] 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.

[40] J. Veness, K. S. Ng, M. Hutter, W. Uther, and D. Silver, "A Monte-Carlo AIXI Approximation," *Journal of Artificial Intelligence Research*, vol. 40, pp. 95–142, 2011.

[41] S. Singh, A. G. Barto, and N. Chentanez, "Intrinsically Motivated Reinforcement Learning," in *Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS)*, Vancouver, B.C., Canada, Dec 2005.

[42] M. Vergassola, E. Villermaux, and B. I. Shraiman, "'Infotaxis' as a strategy for searching without gradients," *Nature*, vol. 445, no. 7126, pp. 406–409, 2007.

[43] N. Ay, N. Bertschinger, R. Der, F. Gttler, and E. Olbrich, "Predictive information and explorative behavior of autonomous robots," *European Journal of Physics: Complex Systems*, 2008.

[44] P.-Y. Oudeyer and F. Kaplan, "How can we define intrinsic motivation?" in *Proceedings of the 8th International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems, Lund University Cognitive Studies*, M. Schlesinger, L. Berthouze, and C. Balkenius, Eds., 2008.

[45] T. Berger, "Living information theory," *IEEE Information Theory Society Newsletter*, vol. 53, no. 1, p. 1, 2003.

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

[47] M. Genesereth and N. Love, "General game playing: Overview of the AAAI competition," *AI Magazine*, vol. 26, pp. 62–72, 2005.

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

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

[50] N. Slonim, "The Information Bottleneck: Theory And Applications," Ph.D. dissertation, The Hebrew University, 2003.

[51] A. Junghanns and J. Schaeffer, "Sokoban: A Case-Study in the Application of Domain Knowledge in General Search Enhancements to Increase Efficiency in Single-Agent Search," *Artificial Intelligence, special issue on search*, 2000.

[52] D. Dor and U. Zwick, "SOKOBAN and other motion planning problems," *Comput. Geom. Theory Appl.*, vol. 13, no. 4, pp. 215–228, 1999.

[53] D. Robles and S. M. Lucas, "A simple tree search method for playing Ms. Pac-Man," in *CIG'09: Proceedings of the 5th International Conference on Computational Intelligence and Games*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 249–255.

[54] J. Schmidhuber, "Driven by Compression Progress: A Simple Principle Explains Essential Aspects of Subjective Beauty, Novelty, Surprise, Interestingness, Attention, Curiosity, Creativity, Art, Science, Music, Jokes," *CoRR*, vol. abs/0812.4360, 2008.

[55] J. Lehman and K. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary computation*, vol. 19, no. 2, pp. 189–223, 2011.

[56] P. Capdepuy, D. Polani, and C. L. Nehaniv, "Constructing the Basic Umwelt of Artificial Agents: An Information-Theoretic Approach," in *Proceedings of the Ninth European Conference on Artificial Life*, ser. LNCS/LNAI, F. Almeida e Costa, L. M. Rocha, E. Costa, I. Harvey, and A. Coutinho, Eds., vol. 4648. Springer, 2007, pp. 375–383.

[57] M. Müller, "Challenges in Monte-Carlo Tree Search," 2010, unpublished. [Online]. Available: http://www.aigamesnetwork.org/_media/main:events:london2010-mcts-challenges.pdf

[58] ——, "Fuego-GB Prototype at the Human machine competition in Barcelona 2010: A Tournament Report and Analysis," University of Alberta, Tech. Rep. TR10-08, 2010.

[59] C. Salge, C. Glackin, and D. Polani, "Approximation of Empowerment in the Continuous Domain," *Advances in Complex Systems*, vol. 16, no. 01n02.

APPENDIX A
SOFT-HORIZON EMPOWERMENT COMPLETE ALGORITHM

The soft-horizon empowerment algorithm consists of two main phases. This appendix presents the complete algorithm. In order to make it somewhat independent and concise, it introduces some notation not used in the main paper.

Phase 1 is not strictly necessary, but acts as a powerful optimization by vastly reducing the number of action sequences that need to be analysed. The main contributions presented in this paper are within phase 2.

A. Setup

1) Define a set $\mathcal{A}$ as the list of all possible single-step actions available to the player.
2) Set $n = 0$. Begin with an empty set containing a list of action sequences, $\mathcal{A}^n$.

B. Phase 1

Phase 1 serves to create a set of action sequences $\mathcal{A}^n$ that, most likely, will reach all possible states within n-steps, but will have very few (0 in a deterministic scenario) redundant sequences. In stochastic scenarios that have heterogeneous noise in the environment it may be that those areas are avoided in preference to staying within more stable states, and in these cases you will find there may be some redundancy in terms of multiple action sequences to the same state.

In a deterministic scenario phase 1 can be entirely skipped; the same optimization can be achieved by selecting at random a single action sequence for each state reachable within n-steps (i.e. for each state s select any single action sequence where $p(s|a_t^n) = 1$).

3) Produce an extended list of action sequences by forming each possible extension for every action sequence in $\mathcal{A}^n$ using every action in $\mathcal{A}$; the number of resultant action sequences should equal $|\mathcal{A}^n| \cdot |\mathcal{A}|$. Replace $\mathcal{A}^n$ with this new list and increment n by 1.
   i. Using the channel/transition table, $p(s_{t+n}|a_t^n)$, note the number of unique states (labelled σ) reachable using the action sequences in $\mathcal{A}^n$, always starting from the current state.
4) Produce $p(a_t^n)$ from $\mathcal{A}^n$, assuming an equi-distribution on $\mathcal{A}^n$. Using this, combined with $p(s_{t+n}|a_t^n)$, as inputs to the Information Bottleneck algorithm (we recommend the implementation at (50), pp. 30, see Appendix B). For G, our groups of actions (labelled T in (50)), we set $|G|$ to be equal to σ, such that the number of groups matches the number of observed states. This will produce a mapping, $p(g|a_t^n)$, which will typically be a hard mapping in game scenarios. Select a random value of a from each G (choosing $argmax_{a_t^n} p(g|a_t^n)$ in cases where it is not a hard mapping). Form a set from these selected values, and use this set to replace $\mathcal{A}^n$.
5) Loop over steps 3 and 4 until n reaches the desired length.

C. Phase 2

Phase 2 extends these base n-step sequences to extended sequences of $n + m$-steps, before collapsing them again such that we retain only a set of n-step sequences which can forecast their own futures in the following m-steps available to them.

6) Produce a list of action sequences, $\mathcal{M}$, by forming every possible m-step sequence of actions from the actions in $\mathcal{A}$.
7) Produce an extended list of action sequences, $\mathcal{A}^{n+m}$, by forming each possible extension for every action sequence in the final value of $\mathcal{A}^n$ using every action sequence in $\mathcal{M}$.
8) Create a channel $p(s_{t+n+m}|a_t^{n+m})$ (where $a_t^{n+m} \in \mathcal{A}^{n+m}$) by sampling from the environmental dynamics for our current state (using the game's transition table). For environments with other players, one can use any approximation of their behaviour available and sample over multiple runs or, lacking that, model them with greedy empowerment maximisation based on a small horizon.
9) Calculate the channel capacity for this channel using the Blahut-Arimoto algorithm, which provides the capacity achieving distribution of action sequences, $p(a_t^{n+m})$.
10) Now collapse $p(s_{t+n+m}|a_t^{n+m})$ to $p(s_{t+n+m}|a_t^n)$ by marginalizing over the equally distributed extension of the action sequences:

$$p(s_{t+n+m}|a_t^n) = \frac{\sum_{a_{t+n}^m} p(s_{t+n+m}|a_t^n, a_{t+n}^m)}{|A_{t+n}^m|}$$

where

$$p(s_{t+n+m}|a_t^{n+m}) \equiv p(s_{t+n+m}|a_t^n, a_{t+n}^m)$$

11) Apply the Information Bottleneck (as in (50), pp. 30, see Appendix B) to reduce this to a mapping of action sequences to strategies, $p(g|a_t^n)$ where G are our groups of action sequences grouped into strategies. Cardinality of G sets how many strategy groups you wish to select.
12) We now need to select a representative action, $a^{(\text{rep})}$, from each group g that maximises approximated future empowerment (and weight this on how well the action represents the strategy, $p(g|a_t^n)$, which is relevant if $p(g|a_t^n)$ is not deterministic):

$$a^{(\text{rep})}\big(p(s_{t+n+m}|a_t^n, a_{t+n}^m), g\big) = \arg\max_{a_t^n} \left(p(g|a_t^n) \cdot \sum_{a_{t+n}^m \in A_{t+n}^m} I(a_t^n, a_{t+n}^m; S_{t+n+m}) \right)$$

Where $a_t^{n+m} \equiv a_t^n, a_{t+n}^m$, and using the capacity achieving distribution of action sequences, $p(a_t^{n+m})$, calculated in step 9 above. Note that the mutual information there requires the full channel, but sums over those parts of the channel with the identical n-steps, so algorithmically it is advisable to calculate and store these mutual information values as whilst doing the channel collapse above.

We can now form a distribution of n-step action sequences from the set of values of $a^{(\text{rep})}$ from each action group; these represent a variety of strategies whilst aiming to maximise future empowerment within those strategies.

APPENDIX B
INFORMATION-BOTTLENECK ALGORITHM

The Information Bottleneck algorithm is a variation of rate-distortion, in which the compressed representation is guided not by a standard rate-distortion function but rather through the concept of relevancy through another variable. We wish to form a compressed representation of the discrete random variable A, denoted by G, but we acknowledge that different choices of distortion would result in different representations but it is likely that we have some understanding of what aspects of A we would like to retain and which could be discarded. The Information Bottleneck method seeks to address this by introducing a third variable, S, which is used to specify the relevant aspects of A.

More formally, we wish to compress $I(G;A)$ while maintaining $I(G;S)$. We are looking to retain the aspects of A which are relevant to S, whilst discarding what else we can. We introduce a Lagrange multiplier, β, which controls the trade-off between these two aspects, meaning we seek a mapping $p(g|a)$ which minimises:

$$\mathcal{L} = I(G;A) - \beta I(G;S)$$

The iterative algorithm we present here is from (50).

Input:

1) Joint distribution $p(s,a)$
2) Trade-off parameter β
3) Cardinality parameter σ and a convergence parameter ϵ

Output:

A mapping $p(t|a)$, where $|G| = \sigma$. For the scenarios in this paper this typically a hard mapping, but it is not necessarily so.

Setup:

Randomly initialise $p(g|a)$, then calculate $p(g)$ and $p(s|g)$ using the corresponding equations below.

Loop:

$$1)\, P^{(m+1)}(g|a) \leftarrow \frac{P^{(m)}(g)}{Z^{(m+1)}(a,\beta)} e^{-\beta D_{KL}[p(s|a)\|p(s|g)]}, \forall\, g \in \mathcal{G}, \forall\, a \in \mathcal{A}.$$

$$2)\, P^{(m+1)}(g) \leftarrow \sum_a p(a) P^{(m+1)}(g|a), \forall\, g \in \mathcal{G}.$$

$$3)\, P^{(m+1)}(s|g) = \frac{1}{P^{(m+1)}(g)} \sum_a P^{(m+1)}(g|a) p(a,s), \forall\, g \in \mathcal{G}, \forall\, s \in \mathcal{S}.$$

Until:

$$JS(P^{(m+1)}(g|a), P^{(m)}(g|a)) \leq \epsilon$$

Where JS is the Jensen-Shannon divergence, based on D, the Kullback–Leibler divergence:

$$JS(P \parallel Q) = \frac{1}{2} D(P \parallel M) + \frac{1}{2} D(Q \parallel M)$$

A. Parameter Values

In the present paper, for all scenarios β = 3, and ϵ = 0.0000001.

The algorithm is based on a random initialisation of $p(t|a)$, so it is usually beneficial to use multiple runs of the algorithm and select that with the best result. Throughout the present paper we used 200 runs of the Information Bottleneck algorithm in each instance it was used.

5.3 Paper 3: Additional Result

5.3.1 Gambler's Problem

An additional scenario had to be excluded from the paper due to space constraints; it is included here prior to a discussion of the paper.

The Gambler's Problem, which was, to our knowledge, first studied in Dubins and Savage (1976), has been addressed several times before, notably in Sutton and Barto (1998) where Reinforcement Learning is used to solve it. The scenario, as presented here, is one in which a gambler, with some current balance of money given by a positive integer, b, with $b < t$ where t is a target balance, must repeatedly make bets until he is either bankrupt or reaches (or exceeds) the target balance, t.

In each betting round, the gambler wagers some positive integer value w where $w \leq b$ and has a probability of winning given by p; here we only examine the case p=0.4 (as in Sutton and Barto (1998)). If the bet is won, then the gambler's balance is increased by the amount wagered, and if the bet is lost, then their balance is decreased accordingly. Thus, the new balance becomes:

$$b' = b + w \text{ with probability } p$$
$$b' = b - w \text{ with probability } 1 - p$$

The objective is to find the optimal betting policy: the best bet possible for all possible balances (i.e. those that maximise the probability of reaching the target balance).

A general property of the problem is that for $p > 0.5$, the gambler will try to prolong the game as much as possible. In our case, where $p < 0.5$, the situation is more interesting, and the optimal strategies (which are not unique in general) will strive to make as few bets

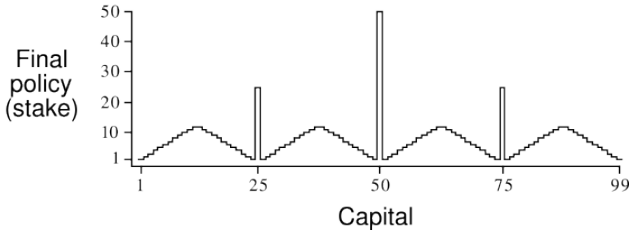


Figure 5.1: *The results for the Gambler's Problem from Sutton and Barto (1998), with p=0.4 and a target balance of 100.*

as possible. This makes sense given that each bet is more likely to lose than to win and so prolonging the betting is a bad proposition.

This problem is an exception to those addressed in Paper 3, as it includes an explicit goal state as an integral part of the scenario. Because of that it works to demonstrate one way in which an explicit goal state, if necessary, can be incorporated into the soft-horizon empowerment formalism.

I handle this by creating a 'paradise' state; a winning bet that carries you past the target balance puts the agent into this paradise state. In this state the agent begins with a balance equal to the target balance and any bets placed whilst having a balance above the target threshold would be a winning bet. This paradise state does not explicitly define higher balances to be better, but by allowing a player to exactly predict the outcome of any move the player's empowerment is increased.

Furthermore, an increased budget provides for an increased set of possible moves (you have more options what to bet), which also increases empowerment. So the attraction to this paradise state comes from the elimination of stochasticity, and the increase in the number of available moves.

Fig. 5.2 shows the results of applying soft-horizon empowerment to the Gambler's Problem using soft-horizon empowerment, where n=1 and m=2. Fig. 5.3 shows the results for n=1, and m=3. The algorithm was run for each starting balance, and then a bandwidth constraint

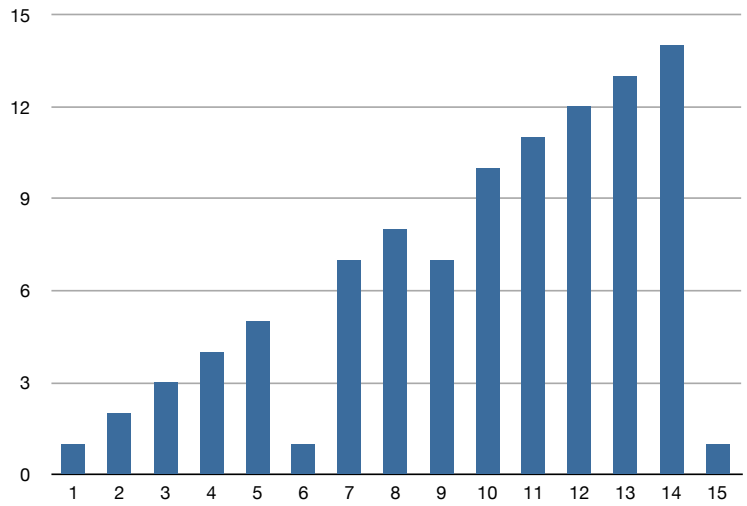


Figure 5.2: *The betting policy produced by soft-horizon empowerment where n=1, m=2, sampled over 100 runs with identical results. The probability to win a bet p=0.4, with a target balance $t = 16$. It can be seen that, in general, the strategy is a naive 'bet it all' policy, but there are points of interest at a starting balance of 6, 9 and 15. The bets at starting balances 9 and 15 represent (non-unique) optimal bets - the smallest single bet that can reach the target balance.*

was applied to select a single action, which would be the action contributing the most to the empowerment. This process produced a betting policy for what amount, w, to bet for each starting balance, b, shown by the graph. In both instances we took 100 runs, and in both instances the results were identical across all the runs.

The channel $p(s_{t+1+m}|a)$ gives a probability of ending in each of the possible states based on an approximation of the capacity achieving distribution across the future m actions that would follow for the bet a.

The target balance was set at $t = 16$. Key states to pay attention to are the *pivot* balance values of 4, 8, and 12; these are points where there is an obvious best bet to make. At 12, it is clear to see you should bet 4, and you either win, or fall back to the next pivot point of 12. At 8, where the player's balance matches exactly the amount it needs yet to win, it seems clear that a single bet of the entire balance is the best strategy (there is no advantageous 'fall back'

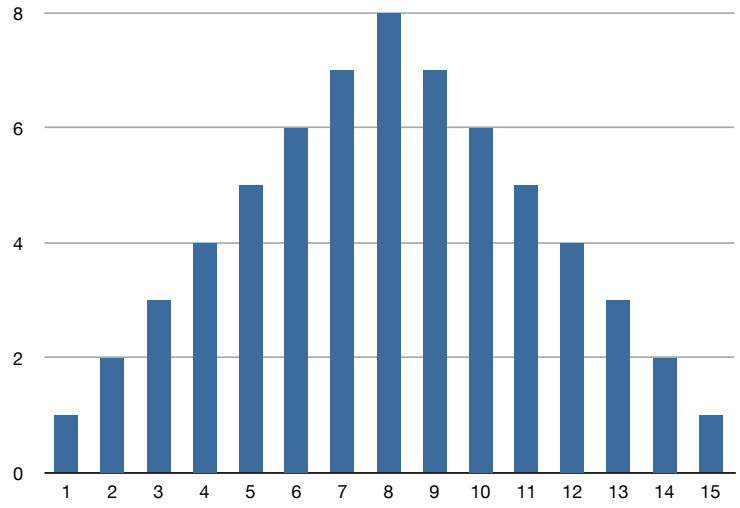


Figure 5.3: *The betting policy produced by soft-horizon empowerment where n=1, m=3, sampled over 100 runs with identical results. The probability to win a bet p=0.4, with a target balance* $t = 16$*. The betting policy equates to the bold play strategy proved optimal in Dubins and Savage (1976).*

position as with the case of 12). The case of 4 can be thought of, in this instances, as a 'sub game' where the player first needs to get to the immediate goal of 8; again there is no fallback and betting our entire budget makes sense and in the case of success, one reaches $b = 8$ with the opportunity to win with the next move.

In particular, for the Reinforcement Learning Gambler's problem solution from Sutton and Barto (1998), it is the pivot points that stand out in the policy space. These results are shown in Fig. 5.1, with a target of $t = 100$.

Returning to empowerment, the results for m=2 in Fig. 5.2 show an inconsistent and sub-optimal betting strategy, with the general trend being the 'bet it all' approach. However, the bets at starting balances 9 and 15 represent (non-unique) optimal bets; they are, in each case, the smallest single bet that can reach the target balance. Below the halfway point, at a balance of 8, one optimal approach is to bet your entire balance which is what is seen for all cases other than a starting balance of 6. In total 9 of the 15 values for w are optimal.

When we increase m to m=3 in Fig. 5.3, we see a dramatic shift in the policy and an improvement in results. The results for m=3 show a betting strategy that equates to the *bold play* policy that was proven to be (non-uniquely) optimal in Dubins and Savage (1976).

These two results show how an insufficient horizon can lead to dramatically different behaviour. In contrast to the changes in horizon observed in the scenario with the blocked door shown in Fig. 8 of the paper, where the change in horizon resulted in a change in a tactical change that was still sensible, here we can see that in some scenarios an inappropriate horizon can simply result in a move away from optimality.

This example shows that the principles of maximising control and 'folding back' future states in order to anticipate and maintain future control extend beyond the realm of deterministic gridworld scenarios and into non-deterministic scenarios in other spaces. It also demonstrates possible strategies to incorporate explicit goals seamlessly into the empowerment framework when this is required.

5.4 Paper 3: Discussion

The paper introduced soft-horizon empowerment which provides a method of anticipating future utility, and a way of using that to generate rich action policies. Furthermore, it encapsulates a novel method of grouping actions into strategies; this *strategy clustering* method is presented within the framework of empowerment but generalises more widely.

The scenarios of 'Pac-Man' and the Gambler's Problem highlight the aspect of anticipating future utility and acting in order to maintain it, and the Sokoban inspired box-pushing scenarios highlight the general method of strategy clustering and worked as a convenient method to compare against standard mobility measures.

5.4.1 Anticipated Utility

The paper introduced soft-horizon empowerment, which introduces a second horizon, with the goal being to embed a method of anticipating future utility into the empowerment framework. This allows empowerment to generate action policies which aim to maintain utility and thus enable empowerment to drive behaviour, rather than solely evaluate states.

From States to Actions

With a second horizon, the channel on the perception-action loop changes from $p(s_{t+n}|a_t^n)$ to $p(s_{t+n+m}|a_t^n)$. Previously the channel ($p(s_{t+n}|a_t^n)$) could be informally described as 'what is the distribution of states I will end up in after n time steps if I spend those time steps enacting the action sequence a_t^n. However, the new channel ($p(s_{t+n+m}|a_t^n)$) changes this to be 'what is the approximate distribution of states I may be in after $n+m$ time steps, if I spent the next n time steps enacting the action sequence a_t^n.

The channel now has baked into it, in a manner consistent with the pre-existing information theoretic scaffolding, an understanding of anticipated future utility, and can now use that to change the capacity achieving distribution of actions.

With this change, empowerment now shifts from state-centric (how many actions available, regardless of the resulting states they reach) to action-centric (what actions are available, such that the resulting states maintain maximum anticipated control). This is an important change as it allows empowerment to drive behaviour in a fashion that is entirely consistent with both the existing framework and with the 'spirit' of empowerment.

Comparison with Greedy Mobility

Another aspect highlighted in the paper, is the comparison of empowerment against naive greedy mobility; previously it was possible to compare empowerment against mobility, but

the move towards action-oriented empowerment now allows also comparing the change in behaviours.

In Section VIII of the paper, I introduced a simple scenario with an agent on a wall which could 'jump' down into either room (shown in in Fig. 13 in the paper).

To compare greedy mobility algorithm to empowerment, each algorithm was to select a single action sequence of 2 steps, with a second horizon of 2 additional steps. The scenario was designed such that most initial 2-step action sequences the agent can perform (stay on the wall, go one step away from it, or stand next to it without moving north/south) lead to a mobility of 9 moves ($\mathfrak{E} = 3.17$). Moving off the wall then north or south results in a mobility of 8 moves ($\mathfrak{E} = 3.00$).

Initially, the experiment was run without any noise, such that it was entirely deterministic. Soft-horizon empowerment selected only from 2 moves (EE & WW) with equal probability, whereas greedy mobility chose from all action sequences approximately equally. In this instance, empowerment chose always to move away from the wall as an artefact of the way it forecasts future empowerment.

Interestingly, when the scenario was run a second time, after the introduction of noise in to one of the two chambers (in the form of a 50% probability that in any given run the actuation of a direction would result in its opposite being picked). This type of noise has no impact on mobility, but does impact the ability for an agent to control what states it visits.

In this instance, empowerment immediately favoured the room that was constantly deterministic, whereas naive mobility only had a slight shift towards favouring that room.

This comparison with mobility is helpful, especially within the context of games, where mobility is established as standard heuristic (Clune, 2011), as it demonstrates the applicability of a general utility, yet also allows us to highlight how such a utility can be more effective. Empowerment and mobility are clearly very aligned, with standard empowerment and mobility being directly equivalent in deterministic games of perfect information.

However, as we have seen, empowerment is more general in that it can not only recognise the presence of noise and adapt accordingly, but it can also quantifiably measure the effect of that noise on the utility score.

Kiting

In the Pac-Man scenario I highlighted a particular behaviour known as *kiting*, which is seen and used in a variety of games (Uriarte and Ontanón, 2012), where a player runs closely ahead of an enemy player such as to 'drag' them along behind them in a predictable fashion.

This behaviour emerged from soft-horizon empowerment, and was initially hard to interpret (I originally thought it was an error and 'debugged' it for sometime!). Further analysis showed it to be a sophisticated behaviour. As explained in the paper, it is a method of reducing the stochasticity of the environment and affords the agent more control over its environment (including the option of suicide at any point, which does increase empowerment).

I was previously unaware of kiting as a term, but had employed the tactic myself in various games. It was encouraging that a recognised tactic, across multiple different games, emerged from the framework in a self-driven fashion; one of the motivations for my work was to try to drive 'human-like' intuitive behaviours.

5.4.2 Clustering by Strategic Affinity (CLUSTA)

In the paper I introduced a novel method of clustering multiple action sequences together into strategies according to their strategic affinity. I introduced strategic affinity entirely within the information-theoretic framework of empowerment, but discussed how I believe it could extend beyond that framework.

In the hopeful anticipation of future work taking this concept further, it would be helpful to standardise the terminology around this concept.

Any approach of clustering actions according to their strategic affinity may be designated as *Clustering by Strategic Affinity (CLUSTA)*, which allows for a flexible definition of strategic affinity.

Strategic affinity should refer to any method which provides a pairwise 'distance' metric between any two action sequences indicating how strategically aligned they are. This distance measure should be based on how much potential overlap there is in the future states that could be reached should each of these action sequences be enacted (in some additional time window).

A broader definition allows for the strategic affinity measure to be either general or more customised to a specific game/scenario (which may be learned and refined during the lifetime of an agent in method parallel to learning game specific heuristics).

I would encourage fully general methods where possible, however, it need not be based upon a search of states into the future and may be heuristically constructed by some analysis of the current state.

This may mean that some games or scenarios may need some context-specific processing. For example, in measuring the overlap of states in Go, it may be that exact duplicate states are too sparse to drive clustering, such that it is necessary to first process states somehow (e.g. by dividing the board into different regions and comparing those).

In Chapter 6, I offer some further discussion around strategic affinity.

Chapter 6

Discussion and Conclusion

Discussion

Imagination is more important than knowledge. Knowledge is limited.

— Einstein (1929)

6.1 Summary

Garry Kasparov has stated that he believ es the reason a human, with access to an ordinary PC and chess engine, can often beat the strongest specialised chess computers is that humans have a better understanding of what moves deserve the most focus (Kasparov, 2017).

Thought of in the tree search terms of a chess computer, essentially humans are better at pruning the game tree from the outset, which serves to highlight the difference in approaches between humans and computers that play chess. Humans have powerful heuristics and do not need to do the vast processing and evaluation of millions of potential action sequences (which they would be poor at anyway) in order to make potent deductions about what moves might be best.

In other games humans, who may have never played the game before, can often make intuitive decisions and outperform computers - a fact highlighted by design in the game of Arimaa (Syed and Syed, 2003). Whilst the Arimaa challenge was beaten in 2015 (Syed, 2015), the principle is still demonstrated still in Stratego. There are obviously some complex processes at work, likely rooted in pattern matching, that allow humans to do this, and those processes are

almost certainly very different than the exhaustive search based methods that drive computer algorithms.

For this dissertation, I was motivated to understand how we might help develop artificial agents generate 'intuitive' action policies. I was specifically interested in methods that seemed to drive biologically plausible behaviours, such that we might move in the direction of a better understanding of how biotic 'heuristics' can outperform traditional computer algorithms.

With supervised learning methods and reinforcement learning methods, a fitness function or reward is necessary for the learning processes to work. Such rewards or fitness functions require that someone explicitly defines the criteria for success. We might find that such an approach means that the possible paths to ultimate success are constrained by the biases of the entity crafting that fitness function. If our learning approach is capable of recognising patterns or structures at either the local or global levels, it may fail to do so due to a feedback mechanism that is incomplete or only locally optimal. Furthermore, for any system that has any sort of self-learning or adaptation, these explicit rewards will limit the degree to which it might adapt, and any imperfect assumptions may be compounded.

Universal utilities, and specifically methods that require no pre-determined goal, provide a starting point whereby the algorithm retains maximum freedom to select actions, unburdened by any biases from their authors.

In selecting a goalless utility method, empowerment stands out as a method that requires no external reward and, being based in information theory, provides a robust framework for understanding its interactions with the environment. In biology, organisms are (in general) primarily motivated by survival, and any 'fitness' function has that as its ultimate goal. Empowerment generalises this to maintaining control of one's own destiny.

With the empowerment framework as the foundation for the work, several important additions have been made, which I summarise here.

6.1.1 Empowerment

Horizon

Requiring a pre-determined knowledge of an appropriate horizon for any task is not only difficult, but is also in direct contradiction to the goal of trying to develop an approach that is more human-like, intuitive and flexible.

Previously, empowerment needed a pre-selected static horizon, which made it more difficult to utilise in various scenarios without some work to tune it to that scenario, which detracts from the generality of the method.

Ultimately, I suggest that horizons should be entirely dynamic, and extended as necessary based upon context, which may mean different aspects of the same scenario are explored to different depths. However, as an initial approach, I developed the concept of iteratively extending the horizon, which has some attractive traits:

- It is akin to how humans solve problems; it can be flexible based upon the amount of available time. The longer you have the deeper you can explore and think about a problem.
- An iterative approach could provide for adaptively adjusting the horizon based on observable features in the world. For example, an agent may be able to identify an affordance in the distance, and be able to estimate the number of time-steps to reach it. It may then select the initial horizon based upon the time steps needed to reach that affordance, essentially asking 'I could go and manipulate that box over there, which will take me 8 time steps to reach, so before I decide, let me think what else I might do in that time'.
- The idea of sub-goals is very related to the idea of operating within certain time constraints, before then considering what might be done from that point.

However, there are also drawbacks to such an approach as presented in this dissertation. The iterative 'extend and compress' approach is not totally dissimilar to a tree search with aggressive pruning, which does not seem to be a biologically grounded approach. Nonetheless, this criticism is based around the necessity to 'sample' the world in order to construct the channel for empowerment to operate, rather than the iterative approach itself.

I later introduced the idea of having a second horizon, which was less precise but allowed for a forecast of what the future may look like for an agent; I discuss this further below.

Actions & Self Motivation

'Traditional' empowerment provided a general utility for comparing the preferability of states with one another in a goalless scenario. In order to calculate the channel capacity, the Blahut-Arimoto algorithm is used to calculate capacity achieving input distribution (the distribution of actions or action sequences).

However, this distribution of actions has two shortcomings:

1. No understanding of how much empowerment is provided by any of those actions.
2. No concept of how empowering (i.e. by providing future empowerment) these actions may be.

Initially, in Paper 2, I presented an approach which addresses the first of these and, by calculating the mutual information between single actions and destination states, can precisely measure how much of a states empowerment is provided by a specific action in both deterministic and stochastic scenarios.

This was used as the basis of a self-motivated method of action selection, by constraining the number of action sequences the agent may retain, and thus forcing it to select a subset of actions. When used to drive action selection, it turns out this is broadly equivalent to selecting

actions according to a combination of the inverse of the redundancy (i.e. action sequences that lead to the same state as others are less empowering) and the inverse of noise (i.e. action sequences with more predictable outcomes).

That empowerment prefers action sequences with less noise is useful, as is the the drive towards unique action sequences which tends to produce exploratory behaviours.

However, the resulting action policies may include multiple action sequences to states that are nearby one another in the state space (be that geographically or otherwise), and furthermore the destination states may not have much (if any) empowerment.

This, of course, relates to the second point above about how we may produce action policies that provide future empowerment, which was addressed by soft-horizon empowerment.

Soft-Horizon Empowerment

In Paper 3, I presented *soft-horizon empowerment*, which introduced the concept of a second horizon. This second horizon fits neatly into the information theoretic framework of empowerment, and provides a method of anticipating what an agent's future empowerment may be following an available action. This anticipated empowerment is not separate from the current empowerment, and instead refines the measure of the agent's current empowerment to account for anticipated future empowerment.

Furthermore, this anticipated future empowerment of actions is encapsulated within the capacity fulfilling action distribution, meaning that the method previously introduced of constraining the retainable actions in order to drive self-motivated actions, takes advantage of this. Previously actions that 'contributed' the most empowerment were those that led to unique states, but in the new model the actions that contribute the most empowerment are those that themselves are most empowering. Compression now leads to selecting action sequences that looks like they maximise future empowerment.

In the previous iterative model, we observed that the actions selected to make up the policy may be very similar, and lead to almost identical (albeit empowering) states. This is a limitation both for selecting a final action, but also for interim action selections when using the iterative approach of selecting and extending action sequences.

Soft-horizon empowerment overcomes this, because the second horizon provides a method of anticipating how 'nearby' to one another the distribution of states would be, following the enactment of any of the available actions. This *strategic affinity* ensures, in the soft-horizon empowerment model, strategic diversity amongst the retained action sequences.

6.1.2 Strategic Affinity and Strategically Diverse Action Repertoires

Soft-horizon empowerment encapsulates the concept of strategic affinity within it. However, the concept is important enough and, I believe, significant enough to warrant attention as a concept on its own. The realisation of strategic affinity in soft-horizon empowerment is only one possibly formulation, and I believe the concept could apply more broadly.

Thinking in strategic terms is more like we find ourselves thinking, and is demonstrated in game playing. People will usually discuss the various higher level options (strategies) before selecting a specific action in order to enact that higher level plan. For example, in chess we might discuss 'taking control of the centre' or 'defending the queen', before deciding on one and then picking a specific action to enact that strategy.

Thus, in order to generate action policies in a similar fashion, an agent must identify the higher level options. Strategic affinity allows us to do that in a very general fashion.

Having an understanding of the higher level strategies is helpful for a number of reasons:

- It can help to ensure the agent does not leave some areas unexplored or inaccessible. An action policy that naively selected a set of the 'best' actions (as determined by anticipated future utility) might select only actions that represent the dominant strategy.

- When dominant or obvious strategies exist, it is useful to be able to think 'outside the box' and strategic affinity can help with this; I discuss this more in the Section 6.2.
- For explorative behaviours it is useful to be able to understand the task space, such that it can be explored most thoroughly. This sort of grouping can help.
- As demonstrated in the Sokoban inspired scenarios, this approach seems to drive intuitive behaviour. In some situations, such as 'pick 4 actions in this grid world', the actions seem to align very well with Schelling points (Schelling, 1960).

If the guiding principle of empowerment is around maximising ones options, then it is noteworthy that maximising the controllable enactment of strategies is born from the same concept, but operates at a higher level.

Most importantly, I believe this sort of strategic identification is a necessary step in mimicking human approaches. Strategic affinity may not be the final solution, but is helpful for guiding research in that direction.

6.2 Future Research and Possible Applications

The work presented in this dissertation provides a novel approach towards generating action policies that produce 'intuitive' behaviours, which are not dissimilar to what we might expect a human to try.

On the one hand, the information theoretic basis allows introspection of the behaviours, and provides a strong framework for improving our understanding. However, on the other hand it is clear that, whilst we are motivated to replicate biologically grounded behaviours, the information theoretic approach is very unlikely to be how such behaviours are driven in nature.

Future research can, therefore, likely be broken into two main threads. The work, as it stands, has possible immediate applications in various areas, such as game-playing, with a view to making the model more viable as a model for deeper applications.

The second approach would be attempting to further extend the framework, and apply it to a broader array of scenarios, with a long-term view of improving biological plausibility. This might help us drive towards more heuristic based models, which would better mimic biology, perhaps be more computationally efficient, and possibly help produce models capable of more complex cognitive tasks.

6.2.1 Possible Applications (Short Term)

General Game Playing

There is a lot of exciting research in game playing at the moment, with the application of ANNs as a layer on top of Monte-Carlo Tree Search methods.

There are compelling reasons to think that the ability to identify strategies, in a general fashion, may prove an interesting method for helping to guide such tree search methods. This may be as a method of pruning the search tree, or could simply apply as a weighting mechanism in a fashion similar to the UCT variation of MCTS.

There is potential that this may improve performance in the short term, whilst also being a step towards moving toward a more human-like approach to considering moves.

In addition, empowerment may provide the basis for an interesting heuristic for general game playing, both as a standalone heuristic for evaluating and comparing states, but also potentially as a method to guide systems that attempt to learn domain-specific heuristics.

Swarm Robotics

Examining multi-agent systems within the framework of empowerment has already been the focus of research (Capdepuy et al., 2007, 2012), with some interesting results.

Swarm robotics could be an area that would potentially benefit from the strategic affinity. Swarms could co-ordinate their efforts by dividing the work between them according to strategies.

How well this might work outside of a simulated environment though is hard to know without further research.

6.2.2 Possible Applications (Long Term)

Medical Applications

There are a couple of aspects of this work which, in the longer term, may provide some value as a tool with medical applications. Prosthetic technologies have made great strides forward with regard to the degree of control that they provide; note that they fit nearly within the perception-action loop formalism. An empowerment-like approach of maximising the channel capacity of future prosthetics is an interesting concept.

A second idea might be as a method of guiding development of such devices, by providing a measure to the degree to which actuation is 'outsourced' to the design of the device. If we consider, again, walkbot (Collins et al., 2001) which was capable of walking down inclines alone, a framework which measures the degree to which such actuation can be baked into the design of prosthetic devices could be helpful.

Search and Rescue

Search and rescue robots, by their very nature, are likely to encounter situations which have not been considered or which are difficult to plan for. In these situations, a method for driving sensible behaviours is obviously useful. In order to use an empowerment-like model, such a robot would the require ability to model the dynamics of its environment, but could then make sensible decisions without needing to await further instructions (which may or may not be possible).

Exploratory Robots

In addition to search and rescue, robots are being used increasingly for exploration. This ranges from drones being used to explore small, hard to reach, areas of terrain through to various crafts and robots involved with interplanetary exploration.

Again, unforeseen circumstances are more likely for robots tasked with these sort of jobs, and methods of making sensible decisions in those cases are key to survival. In addition, there is scope that being able to evaluate the various options and group them into broad strategies may help, both with fully automated decision making and also when working in tandem with human controllers.

6.2.3 Future Research

Strategic Affinity

Strategic Affinity is a key result from this dissertation, and is an aspect of the work that might also be applicable outside of the context of empowerment.

However, like other aspects of the work, it produces intuitive and understandable results, but the mechanism by which it arrives at them is unlikely to be how we might do so. The

concept of the overlapping futures is a coherent with our view of what a strategy is; however, calculating this overlap using any sort of well defined distribution over future states is not coherent with how we might imagine ourselves forming strategies.

An interesting line of research would be applying pattern-recognition to the task. It would likely need to be somewhat context specific, but could also be general, and may provide an alternative approach to approximating a strategic affinity like measure.

One possible approach, which I frame here in the context of games, would be to 'blur' the playing board and compare it with other blurred states. In Go, for example, this would equate to understanding the broad territory distribution across the board.

An alternative approach may be simply to break a game into sub-games and evaluate a change of state in any of them. In Go, again, this could be to simply consider separate areas of the board in isolation.

I believe that not only would such work help us understand new facets of intelligence, but could also have immediate applications in games playing (see applications section above).

Finally, there is an opportunity for strategic clustering methods to drive 'out of the box' thinking. Sometimes a strategy is so dominant that most of the high utility actions are all variations of it, which means that alternative approaches may not be considered. This is limiting in a variety of ways, especially in dynamic environments or scenarios with an antagonist, where the dominant strategy could become inviable, and having alternative options is essential.

Heuristics

As demonstrated by the parallels between Infotaxis and the search behaviours of moths, it is clear that nature has an extremely potent ability to develop heuristics. Biological organisms are able, with astounding ease, to spot salient features of environment and can also act instinctively in the face of (mathematically) complicated tasks.

For example, a child can catch a ball with relative ease, whereas a computer needs the equations for calculating the parabola of flight.

The empowerment framework is a good candidate for helping guide research into developing general heuristics. That may be as a reward system applied to another adaptive learning approach, or possibly as a tool to compare against.

If we were able to improve our heuristics generation, the applications for artificial agents would be plentiful.

Context Recognition

Most multi-cellular organisms, as well as most artificial agents, find themselves in various different scenarios, which demand different behaviours. This can range from changes in internal state, such as homeostatic drives, or from the external state of the world, such as location within the environment.

Many areas of the work in this dissertation are related to this concept of 'contexts':

- It could be advantageous to introduce the concept of 'directed empowerment' which relates to the specific goals of an agent. Thus empowerment could change depending on context, which may mean only relevant parts of the sensorimotor apparatus are used, and thus different behaviours come about.
- Different contexts could indicate different default horizons are appropriate, which could mean an agent is more efficient in its processing (an evolutionary advantage).
- A context may indicate different sub-goals, which could mean that only some identified strategies are relevant and need to be considered. This would be especially useful for homeostatic contexts, and where there is a metabolic cost to assessing options.

Bibliography

Bibliography

Anthony, T., Polani, D., and Nehaniv, C. L. (2008). On preferred states of agents: how global structure is reflected in local structure. In Bullock, S., Noble, J., Watson, R., and Bedau, M. A., editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 25–32. MIT Press, Cambridge, MA.

Anthony, T., Polani, D., and Nehaniv, C. L. (2011). Impoverished empowerment: 'meaningful' action sequence generation through bandwidth limitation. In Kampis, G., Karsai, I., and Szathmáry, E., editors, *Artificial Life. Darwin Meets von Neumann. ECAL 2009. Lecture Notes in Computer Science*, volume 5778. Springer.

Anthony, T., Polani, D., and Nehaniv, C. L. (2014). General self-motivation and strategy identification: Case studies based on Sokoban and Pac-Man. *Computational Intelligence and AI in Games, IEEE Transactions on*, 6(1):1–17.

Aristotle (1910). Historia animalium. *The Works of Aristotle. Clarendon Press, Oxford*, pages 486a–633a. Translated by D. W. Thompson.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing. *Network: Computation in Neural Systems*, 3(2):213–251.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193.

Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008). Predictive information and explorative behavior of autonomous robots. *European Physical Journal B*, pages 329–339.

Barabási, A.-L. (2003). *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume Books.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barlow, H. B. (1959). Possible principles underlying the transformations of sensory messages. In Rosenblith, W. A., editor, *Sensory Communication: Contributions to the Symposium on Principles of Sensory Communication*, pages 217–234. The M.I.T. Press.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253.

Bauer, T., Desender, K., Morwinsky, T., and Betz, O. (1998). Eye morphology reflects habitat demands in three closely related ground beetle species (coleoptera: Carabidae). *Journal of Zoology*, 245(4):467–472.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Berger, T. (2003). Living information theory. *IEEE Information Theory Society Newsletter*, 53(1):1.

Bernoulli, D. ([trans. 1954] 1738). Specimen theoriae novae de mensura sortis [exposition of a new theory on the measurement of risk]. *Econometrica: Journal of the Econometric Society*, pages 23–36. Translated by Louise Sommer.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comp.*, 13(11):2409–2463.

Bjornsson, Y. and Finnsson, H. (2009). Cadiaplayer: A simulation-based general game player. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(1):4–15.

Blahut, R. (1972). Computation of channel capacity and rate distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473.

Callaway, E. (2017). Spider gene study reveals tangled evolution. `http://www.nature.com/news/spider-gene-study-reveals-tangled-evolution-1.15578`. Accessed: 2017-04-19.

Campbell, M., Hoane, A. J., and Hsu, F.-h. (2002). Deep Blue. *Artificial intelligence*, 134(1-2):57–83.

Capdepuy, P., Polani, D., and Nehaniv, C. L. (2007). Maximization of potential information flow as a universal utility for collective behaviour. In *Proceedings of the First IEEE Symposium on Artificial Life*, pages 207–213. IEEE Press.

Capdepuy, P., Polani, D., and Nehaniv, C. L. (2012). Perception-action loops of multiple agents: Informational aspects and the impact of coordination. 131(3):149–159.

Castillo, C. (2004). *Effective Web Crawling*. PhD thesis, University of Chile.

Chaitin, G. J. (1982). Gödel's theorem and information. *International Journal of Theoretical Physics*, 21(12):941–954.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.

Clark, D. and Sokoloff, L. (1999). Circulation and energy metabolism of the brain. *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, pages 637–669.

Clune, J. E. (2011). Heuristic evaluation functions for general game playing. *KI - Künstliche Intelligenz*, 25(1):73–74.

Collins, S. H., Wisse, M., and Ruina, A. (2001). A three-dimensional passive-dynamic walking robot with two legs and knees. *The International Journal of Robotics Research*, 20(7):607–615.

Coulom, R. (2007). Monte-Carlo tree search in Crazy Stone. In *Proceedings of Game Programming Workshop*, pages 74–75.

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

Darwin, C. (1859). *On the origin of the species by natural selection*. Murray.

Der, R., Güttler, F., and Ay, N. (2008). Predictive information and emergent cooperativity in a chain of mobile robots. In *Artificial Life XI*, pages 166–172. MIT Press.

Der, R. and Martius, G. (2011). In *The playful machine: theoretical foundation and practical realization of self-organizing robots*, pages 107–145. Springer-Verlag Berlin Heidelberg.

Der, R., Steinmetz, U., and Pasemann, F. (1999). Homeokinesis - a new principle to back up evolution with learning. In Mohammadian, M., editor, *Computational Intelligence for Modelling, Control, and Automation*, volume 55 of *Concurrent Systems Engineering Series*, pages 43–47. IOS Press.

Dowe, D. and Hajek, A. (1997). A computational extension to the turing test. In *Proceedings of the 4th Conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*, volume 1.

Dowe, D. L., Hernández-Orallo, J., and Das, P. K. (2011). Compression and intelligence: social environments and communication. In *Proceedings of the 4th international conference on Artificial general intelligence*, pages 204–211. Springer-Verlag.

Du, F., Zhu, X.-H., Zhang, Y., Friedman, M., Zhang, N., Uğurbil, K., and Chen, W. (2008). Tightly coupled brain activity and cerebral atp metabolic rate. *Proceedings of the National Academy of Sciences*, 105(17):6409–6414.

Dubins, L. E. and Savage, L. J. (1976). *Inequalities for Stochastic Processes: How to Gamble If You Must*. Dover, New York.

Einstein, A. (1929). *The Saturday Evening Post*. Saturday Evening Post Company.

Elton, C. S. (1927). *Animal ecology*. University of Chicago Press.

Enzenberger, M., Muller, M., Arneson, B., and Segal, R. (2010). Fuego—an open-source framework for board games and go engine based on monte carlo tree search. *IEEE Transactions on Computational Intelligence and AI in Games*, 2(4):259–270.

Eves, H. W. (1990). In *An introduction to the history of mathematics*, page 427. Brooks/Cole.

Feynman, R. (1965). *The Character of Physical Law*. Messenger lectures on the evolution of civilization. BBC.

Genesereth, M. and Love, N. (2005). General game playing: Overview of the AAAI competition. *AI Magazine*, 26(2):62–72.

Gibson, J. J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, Acting and Knowing: Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Associates.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton, Mifflin and Company.

Grinnell, J. (1917). The niche-relationships of the California Thrasher. *The Auk*, 34(4):427–433.

Hamming, R. (1991). *The Art of Probability: For Scientists and Engineers*. Advanced Book Classics. Avalon Publishing.

Hernández-Orallo, J. and Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539.

Hernández-Orallo, J. and Minaya-Collado, N. (1998). A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of International Symposium of Engineering of Intelligent Systems (EIS 1998)*, pages 146–163. ICSC Press.

Hutchinson, G. E. (1957). Concluding remarks. *Cold spring harbor symposium on quantitative biology*, 22:415–427.

Hutter, M. (2001). Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions. In De Raedt, L. and Flach, P., editors, *Machine Learning: ECML 2001*, volume 2167 of *Lecture Notes in Computer Science*, pages 226–238. Springer Berlin / Heidelberg.

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin.

Hutter, M. (2006). Human knowledge compression contest: Frequently asked questions & answers. `http://prize.hutter1.net/hfaq.htm`. Accessed: 2017-04-24.

Janis, C. M. and Thomason, J. (1995). Correlations between craniodental morphology and feeding behavior in ungulates: reciprocal illumination between living and fossil taxa. *Functional morphology in vertebrate paleontology*, pages 76–98.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Jung, T., Polani, D., and Stone, P. (2011). Empowerment for continuous agent-environment systems. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 19(1):16–39.

Kappelman, J. (1988). Morphology and locomotor adaptations of the bovid femur in relation to habitat. *Journal of Morphology*, 198(1):119–130.

Kasparov, G. (2010). *How Life Imitates Chess: Making the Right Moves, from the Board to the Boardroom*. Bloomsbury Publishing.

Kasparov, G. (2017). *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins*. Hachette UK.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004a). Organization of the information flow in the perception-action loop of evolved agents. In Zebulum, R. S., Gwaltney, D., Hornby, G., Keymeulen, D., Lohn, J., and Stoica, A., editors, *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, pages 177–180. IEEE Computer Society.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004b). Tracking information flow through the environment: Simple cases of stigmergy. In *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 563–568. The MIT Press.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005a). All else being equal be empowered. In Capcarrère, M. S., Freitas, A. A., Bentley, P. J., Johnson, C. G., and Timmis, J., editors, *Advances in Artificial Life: Proceedings of the 8th European Conference on Artificial Life*, volume 3630 of *Lecture Notes in Artificial Intelligence*, pages 744–753. Springer.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005b). Empowerment: A universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, volume 1, pages 128–135. IEEE Press.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2008). Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12):e4018.

Kwak, J.-S. and Kim, T.-W. (2010). A review of adhesion and friction models for gecko feet. *International Journal of Precision Engineering and Manufacturing*, 11(1):171–186.

Laughlin, S. B., De Ruyter Van Steveninck, R. R., and Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, 1(1):36–41.

Mahoney, M. (2006). comp.compression mailing list: Introducing the Hutter Prize for lossless compression of human knowledge. `https://groups.google.com/forum/#!topic/comp.compression/Pwlq6pkyc8s`. Accessed: 2017-04-24.

Mahoney, M. V. (1999). Text compression as a test for artificial intelligence. In *Proceedings of the National Conference on Artificial Intelligence, AAAI, John Wiley & Sons Ltd*, pages 486–502.

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Mehat, J. and Cazenave, T. (2008). Monte-Carlo tree search for general game playing. Technical report, Dept. Informatique, Université Paris.

Ortega, D. and Braun, P. (2011). Information, utility and bounded rationality. *Artificial general intelligence*, pages 269–274.

Pfeifer, R. and Bongard, J. C. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence (Bradford Books)*. The MIT Press.

Pfeifer, R. and Gómez, G. (2009). Morphological computation – connecting brain, body, and environment. In Sendhoff, B., Körner, E., Sporns, O., Ritter, H., and Doya, K., editors, *Creating Brain-Like Intelligence*, pages 66–83. Springer-Verlag.

Polani, D. (2011). An informational perspective on how the embodiment can relieve cognitive burden. In *Proc. IEEE Symposium Series in Computational Intelligence 2011 — Symposium on Artificial Life*, pages 78–85. IEEE.

Prokopenko, M., Gerasimov, V., and Tanev, I. (2006). Evolving spatiotemporal coordination in a modular robotic system. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., and Parisi, D., editors, *From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy, September 25-29 2006*, volume 4095 of *Lecture Notes in Computer Science*, pages 558–569. Springer.

Reed, E. S. (1996). *Encountering the world: Toward an ecological psychology.* Oxford University Press.

Rétaux, S. and Casane, D. (2013). Evolution of eye development in the darkness of caves: adaptation, drift, or both? *EvoDevo*, 4(1):26.

Ryabko, B. and Reznikova, Z. (1996). Using Shannon entropy and Kolmogorov complexity to study the communicative system and cognitive capacities in ants. *Complexity*, 2(2):37–42.

Schelling, T. C. (1960). The strategy of conflict. *Cambridge, Mass*.

Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4:234–242.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Simon, H. A. (1957). *Models of man: social and rational; mathematical essays on rational human behavior in a social setting*. Wiley, New York.

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

Świechowski, M. and Mańdziuk, J. (2015). Specialized vs. multi-game approaches to AI in games. In *Intelligent Systems' 2014*, pages 243–254. Springer.

Syed, O. (2015). The Arimaa challenge. `http://arimaa.com/arimaa/challenge/2015/`. Accessed: 2018-03-10.

Syed, O. and Syed, A. (2003). Arimaa - a new game designed to be difficult for computers. *International Computer Games Association Journal*, 26:138–139.

Tisdell, C. A. (1996). *Bounded rationality and economic evolution : a contribution to decision making, economics, and management*. Edward Elgar, Cheltenham, UK.

Touchette, H. and Lloyd, S. (2004). Information-theoretic approach to the study of control systems. *Physica A: Statistical Mechanics and its Applications*, 331(1):140–172.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):460.

Uriarte, A. and Ontanón, S. (2012). Kiting in RTS games using influence maps. In *Eighth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Varela, F. J., Thompson, E. T., and Rosch, E. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, new edition.

Vergassola, M., Villermaux, E., and Shraiman, B. I. (2007). 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445(7126):406–409.

Von Foerster, H. (2003). Disorder/order: discovery or invention? In *Understanding Understanding*, pages 273–282. Springer.

von Goethe, J. W. (1870). *Goethe's Spruche in Prosa: Zum ersten Mal erläutert und auf ihre Quellen zurückgeführt von G. von Loeper*. Gust. Hempel.

Weiss, M. D. (1987). *Conceptual Foundations of Risk Theory*. Technical bulletin (United States. Department of Agriculture). U.S. Department of Agriculture, Economic Research Service.

Williamson, O. E. (1981). The economics of organization: The transaction cost approach. *The American Journal of Sociology*, 87(3):548–577.