UNIVERSITÄT OSNABRÜCK

Fachbereich Humanwissenschaften
Institute of Cognitive Science

Bachelor's Program Cognitive Science

Exposé

Towards a minimalistic free energy agent

Esther Chevalier
Matriculation number 972437
12/10/2020

Version 1

Esther Chevalier
Springmannskamp 3
49090 Osnabrück

E-mail: echevalier@uos.de
Tel: 01573 57 20 346

Contents

1 Motivation **1**

1.1 Current state of research 1

1.2 Problem 1

1.3 Research question 2

2 Suggested methodology **3**

2.1 Thesis goal and objectives 3

2.2 Methods 3

2.3 Preliminary thesis structure 4

3 Organisational matters **5**

3.1 Preliminary time schedule 5

3.2 Relevant literature 6

3.3 Further remarks 6

4 Bibliography **7**

1 Motivation

1.1 Current state of research

Karl Friston proposes and defends a new formulation of habit-forming, learning and behavior optimization in biological agents [1]. The principle he developed over the years rests upon a theoretical framework previously used in statistical thermodynamics. In particular, the concept developed by Richard Feynman, the variational free energy is central to Friston's hypothesis about behavior learning. The free energy framework has already been used to formulate and model research questions coming from multiple disciplines, such as computational psychiatry, theoretical biology and artificial intelligence [2]. Friston hypothesize the free energy principle as being more than a model of an agent's behavior. Instead, he considers it to be the only strictly necessary feature for determining if a biological system is an agent [3].

The free energy principle has a range of properties making it to an elegant model of an agent's behavior and learning patterns. In particular, learning processes in an artificial agent could be described without using external rewards and penalties, unlike reinforcement learning. The behavior of the agent can be described as "Act to see what you expect to see". The optimization of a single value, the free energy of each action taken by the agent, leads the agent to make assumptions about the environment and to act optimally given its assumptions and its perceptions. The emerging behavior is characterized by the inferences made by the agent and driven by the action it takes - it is therefore called active inference. Moreover, the agent can optimize the inferences it makes by learning progressively. The agent's model about the environment's causal structure and other parameters are optimized incrementally.

1.2 Problem

The free energy principle stays obscure and difficult to understand [4], due amongst other things to some inconsistencies in Friston's argumentation and the formulation

of the free energy principle [5]. The topic remains primarily investigated by Friston himself, as first or second author. There is a lack of a comprehensive review for early-career scientists who do not have a formal training in physics or other disciplines necessary to understand the free energy framework in its current formulation. The difficulty understanding the topic is worsened by the unavailability of source code used in free energy agent simulation in Friston's work e.g. [6].

In addition, the free energy principle roots in a deeply physical framework, visible by the vocabulary and concepts used for its formulation [1]. The free energy has since then been adapted to an information theoric framework [2]. This reformulation begs the question of which concepts and tools are essential for a formulation of the free energy principle and which are not. The free energy principle could be then explained with the lighter version, using the strictly necessary concepts and ideas. This formulation could, if done properly, be more accessible to students and scientists.

1.3 Research question

Currently, the research question which will guide this bachelor thesis is as follows: Can the free-energy framework be adapted to a simple agent, with (relatively) few features? If so, which features are strictly necessary in order to make the agent act and learn efficiently?

The research question is subject to change during the work process.

2 Suggested methodology

2.1 Thesis goal and objectives

The goal of this thesis is to implement a free energy agent capable of making inferences in a given environment and learning new behavior based on those inferences. The agent should be as minimalistic as possible. This allows to understand which features an agent strictly needs in order to optimize its free energy and therefore its behavior. This kind of work has not been done in the current literature yet, although there is at least one (non-published) article which pursue this exact goal [4].

The thesis aims to explain the free-energy principle as concisely as possible to avoid distraction from its core mechanism. It should provide interested students with basic knowledge in probability and information theory the necessary tools to understand active inference and the associated learning processes. The previously mentioned implementation will be used as a concrete example. The challenge lies in narrowing down the free-energy principle to its essence. During the process, it will become clear if the free-energy principle is an adequate tool to make a simple agent behave coherently in a given environment. It will also be part of the project to analyze the weaknesses and benefits of the free-energy approach in a simple context.

2.2 Methods

The thesis will include a theoretical and practical section. The theoretical section will primarily rely on a literature research. It aims to lay the foundation of theoretical tools needed to understand the free-energy principle. The hands-on work is about writing a relatively short piece of code, doing both active inference and learning to use the free-energy framework. The piece of code provides a sandbox for all interested readers and a tool for understanding the free-energy principle with a hands-on experiment.

The code will be written using Python 3.7 and various packages, e.g. NumPy, matplotlib and others. The code will presumably include a small neural network, used for the learning process. If used, it will be implemented using the library TensorFlow. The code will be available on GitHub.

Additionally, the thesis will include a section analyzing the fitness of the framework for this particular agent given its goal and its environment. Strengths and weaknesses of this framework will emerge from the upcoming coding work process and the related literature research. It is not part of the thesis to make an explicit comparison to other learning or inference methods (such as reinforcement learning or Bayesian inference)

2.3 Preliminary thesis structure

The thesis will contain three chapters; the theoretical background, the practical part and the analysis of the resulting agent. The theoretical background should cover every aspect necessary to understand the framework. It is still not clear, which information is essential for the reader to understand the free-energy principle as this will emerge from the coding process.

The practical part will explain and go through the code in a step-by-step manner. The final section will contain an analysis of the previous code, it's main characteristics, how it compares to Friston's previous work and its strengths and weaknesses.

1. Theoretical background
2. Implementation of a minimalistic agent
3. Strengths and weaknesses of the framework

3 Organisational matters

3.1 Preliminary time schedule

	Duration	Start date	End date	Size	Remarks
Practical part - Programming of agent	3 weeks	12 Oct	30 Oct		
Writing - Practical part	2 weeks	26 Oct	6 Nov	6-8 pages	
Registration	1 day	26 Oct	30 Oct		Depends if Prof Pipa is available as second supervisor
Final thesis structure	5 days	9 Nov	13 Nov		
Writing - Theoretical part	3 weeks	16 Nov	4 Dec	10 pages	
Writing - Analytical part	2 weeks	7 Dec	18 Dec	∼7 pages	
Short presentation - (Seminar Academic Writing)	1 day	11 Dec			not obligatory
Winter break	2 weeks	19 Dec	2 Jan		
Writing - Analytical part	1 week	4 Jan	8 Jan	3-4 pages	
Writing - Introduction	2 days	11 Jan	12 Jan	2-3 pages	
Writing - Conclusion	2 days	13 Jan	14 Jan	3-4 pages	
Formatting / Proof-reading	2 weeks	11 Jan	22 Jan		
Printing	1 day	25 Jan	29 Jan		
Submission	1 day	25 Jan	29 Jan		Date not set yet

3.2 Relevant literature

The relevant literature can be found on Mendeley, in the public group free_energy. It accessible via the following URL: `https://www.mendeley.com/community/free-energy-5/`. The content will be updated regularly.

3.3 Further remarks

This exposé is subject to change at any given time, depending on insights coming from the work process and literature research. It will be updated regularly.

4 Bibliography

1. Friston, K., Kilner, J. & Harrison, L. A free energy principle for the brain. *Journal of Physiology Paris* **100,** 70–87. doi:10.1016/j.jphysparis.2006.10.001 (2006).

2. Friston, K. *et al.* Active inference and learning. *Neuroscience and Biobehavioral Reviews* **68,** 862–879. doi:10.1016/j.neubiorev.2016.06.022 (2016).

3. Friston, K. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11,** 127–138. doi:10.1038/nrn2787 (2010).

4. McGregor, S., Baltieri, M. & Buckley, C. L. A Minimal Active Inference Agent, 1–19 (2015).

5. Sims, A. A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology* **29,** 967–980. doi:10.1080/09515089.2016.1200024 (2016).

6. Cullen, M., Davey, B., Friston, K. J. & Moran, R. J. Active Inference in OpenAI Gym: A Paradigm for Computational Investigations Into Psychiatric Illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **3,** 809–818. doi:10.1016/j.bpsc.2018.06.010 (2018).