Does Empowerment Maximisation Allow for Enactive Artificial Agents?

Christian Guckelsberger[1] and Christoph Salge[2]

[1]Computational Creativity Group, Goldsmiths, University of London, UK
[2]Adaptive Systems Research Group, University of Hertfordshire, UK
c.guckelsberger@gold.ac.uk

Abstract

The enactive AI framework wants to overcome the sense-making limitations of embodied AI by drawing on the bio-systemic foundations of enactive cognitive science. While embodied AI tries to ground meaning in sensorimotor interaction, enactive AI adds further requirements by grounding sensorimotor interaction in autonomous agency. At the core of this shift is the requirement for a truly intrinsic value function. We suggest that empowerment, an information-theoretic quantity based on an agent's embodiment, represents such a function. We highlight the role of empowerment maximisation in satisfying the requirements of enactive AI, i.e. establishing constitutive autonomy and adaptivity, in detail. We then argue that empowerment, grounded in a precarious existence, allows an agent to enact a world based on the relevance of environmental features in respect to its own identity.

Introduction

Enactive Artificial Intelligence (AI), as proposed by Froese and Ziemke (2009), represents a framework for the design and evaluation of artificial agents with the goal of fostering intentional agency and sense-making. It results from a critique of the embodied approach to AI (Pfeifer et al., 2005), which was first embraced by Brooks (1991) to overcome several hard problems of *good old-fashioned AI* related to sense-making, in particular the symbol grounding and frame problems. At the centre of Froese and Ziemke's critique is the fact that embodied AI allows for an agent's value function to be externally defined and controlled, which counteracts genuine intentional agency. Inspired by the bio-systemic foundations of enactive cognitive science, they require agents to be genuinely intrinsically motivated in order to afford constitutive autonomy and adaptivity. The goal of this paper is to evaluate whether empowerment maximisation (Klyubin et al., 2008), a bio-inspired, information-theoretic candidate for intrinsic motivation, is sufficient for the realisation of enactive agents with intentional agency.

We first present an overview of embodied AI and of how enactive AI wants to overcome it's shortcomings. This is followed by an in introduction to empowerment, and an in-depth investigation of its role in constitutive autonomy and adaptivity. Crucially, we do not analyse whether enactive AI's requirements are *sufficient* for intentionality and sense-making in artificial agents, but investigate whether they can be met by means of empowerment maximisation.

Criticising Embodied Artificial Intelligence

Situated and embodied cognition together with Enactivism represent three strongly interlinked theories in cognitive science. Situated cognition suggests that cognitive processes emerge from the interaction of an organism and its world, and are thus inseparable from action. Embodied cognition as defined by Rosch et al. (1992) emphasises the role of an agent's physical body in shaping cognitive processes. Given that an agent's body necessarily exists in some place, embodied cognition presupposes situatedness. The theories of embodied situated cognition are supported by a growing body of empirical evidence, highlighting how constraining bodily abilities of human participants can affect e.g. judgement and comprehension processes (cf. Strack et al., 1988; Gallagher, 2005; Havas et al., 2010). The theories are also supported by research in morphological computation (Zahedi and Ay, 2013) and exemplified by "brainless robots" (Pfeifer et al., 2005), which perform otherwise computationally extensive tasks such as walking only by means of their bodily properties, e.g. the constraints and interplay of joints.

Brooks (1991) was the first to bring the ideas from embodied situated cognition to AI research. Since then, embodied AI has developed into a mature framework for modelling artificial agents (cf. Pfeifer and Scheier, 2001), which stands in opposition to good old-fashioned AI with its emphasis on the explicit manipulation of internal symbolic representations. We will outline the embodied approach to AI by reference to a selection of design principles suggested and argued for by Pfeifer et al. (2005). They were split into groups concerning the general design philosophy (P) and the actual design methodology (A).

Embodied AI aims at gaining new insights in the general science of life and mind, as opposed to applied engineering (P-1). Principle P-2 calls for a reduced designer's influence in order to create systems with emergent behaviour.

In the course of this, the designer will face a trade-off between robustness and flexibility of the system (P-3). Principle P-5 adheres to the general theory in stating, amongst other things, that observed behaviour is neither reducible to an agent nor to its environment, and that seemingly complex behaviour can be triggered by simple internal mechanisms. This is where embodied AI differs most from its traditional counterpart. In actual design, agents should never be created in isolation, but with their environment in mind (A-1). Principle A-2 suggests that the proper study of intelligence requires a holistic perspective on agents instead of looking at sub-components only. Principle A-3 states that natural intelligence does not come from algorithms in a central controller, but through the organisation of an agent's sensorimotor loop. Consequently, A-4 says that cognition can be best understood as "appropriate sensorimotor coordination" (Froese and Ziemke, 2009). Finally, the value principle (A-8) requires the agent to be supplied with information about whether a certain action was good or bad, in order to motivate its behaviour.

It is tempting to believe that embodied AI makes one of the biggest challenges of classic AI, the *frame problem*, obsolete. Wheeler (2005) defines it as:

> Given a dynamically changing world, how is a nonmagical system (...) to take account of those state changes in that world (...) that matter, and those unchanged states in that world that matter, while ignoring those that do not? And how is that system to retrieve and (if necessary) to revise, out of all the beliefs that it possesses, just those beliefs that are relevant in some particular context of action? (Wheeler, 2005)

Embodied and situated agents seem to resolve this problem practically: by grounding cognition into their situatedness in a continuously changing world, they do not need to refer to any internal representations. Nevertheless, Froese and Ziemke point out that the presence of a closed sensorimotor loop only addresses the first part of the definition; what is missing is an agent's own capacity to assign relevance to features of the world. They particularly criticise the value principle of embodied AI (A-8), which does not preclude the external assignment of such values or more general goals. More explicitly, they argue that the meaning problem cannot be resolved by injecting values externally, and criticise embodied AI for not demanding an intrinsic perspective. It is a part of our goal to propose a practical solution to this challenge.

The Enactive Approach to Artificial Intelligence

The enactive approach to AI consequently roots in the question of how a system can be designed in which "relevant features of the world show up as significant from the system perspective itself, rather than only in the perspective of the human designer or observer" (Froese and Ziemke, 2009). Froese and Ziemke borrow ideas from enactive cognitive science, a theoretical framework which claims that cognition is embodied, situated and grounded in practical activity. At its core is the idea that individuals do not passively create internal representations of a pre-given external world (Stewart, 2010); instead, they actively generate meaning by constructing their *Umwelt* (Von Uexküll, 1982), i.e. their very own world of significance, through interaction with the environment. According to Rosch et al. (1992), features of the world are not independently out there, but *enacted* through an agent's activity.

Similarly, Froese and Ziemke argue teleologically that behaviour can only be purposeful if it is significant from the system's own perspective. To distinguish simple matter and most artificial agents which are incapable of such intrinsic concern from actual living beings, enactive cognitive science draws on Jonas' notions of *being by being*, as opposed to *being by doing* (Jonas, 1982): while artificial systems can exist without actually doing anything, living systems establish their systemic identity in reaction to the constant threat of becoming a non-being. The latter thus have a precarious existence, which is continually challenged by material or energetic requirements. In order to react to threats, they must be able to assign significance to features of the world.

Jonas suggests that this precarious existence is biologically rooted in an individual's self-organisation, as captured by the concept of *autopoiesis*. Introduced by Maturana and Varela (1987), autopoiesis represents a basic mode of identity. The term only applies to physiochemical systems, and is generalised by the notion of *organisational closure*. A system implementing organisational closure is understood as a network of processes that generate and sustain its identity under precarious conditions, and that form a unity in a containing domain. In their first design principle for fully enactive agents, Froese and Ziemke thus claim that intrinsic teleology requires organisational closure, or in other words, *constitutive autonomy*:

> **EAI-1** (Constitutive autonomy): the system must be capable of generating its own systemic identity at some level of description. (Froese and Ziemke, 2009)

This intrinsic perspective represents the enactive version of embodied AI's value function principle (A-8). In contrast to the synthetic methodology of the embodied approach, it requires the designer to establish the environmental conditions that allow for the emergence of a self-constituting system without direct design influence on the agent architecture.

Although this principle affords a binary significance mechanism, Froese and Ziemke argue that it is not sufficient for sense-making as the enaction of an Umwelt, i.e. as the continuous evaluation of events in relation to maintaining the system's identity. In order to enable an agent to improve its situation or to compensate for some encountered event, it

must be able to distinguish external events more gradually in terms of how they could affect its internal organisation. In other words, they require an agent's Umwelt to not be merely black and white. The capacity to distinguish different tendencies towards non-existence, and to act on them in order to move away from a precarious situation is covered by the concept of *adaptivity* as defined in (Di Paolo, 2005). Additionally, an adaptive agent must be able to act upon its environment to prevent such precarious events in the future. The necessity of adaptivity for sense-making is covered by the second enactive design principle:

> **EAI-2** (Adaptivity): the system must have the capacity to actively regulate its ongoing sensorimotor interaction in relation to a viability constraint. (Froese and Ziemke, 2009)

This viability constraint can either be defined externally or be intrinsically related to the system's identity. Nevertheless, an external viability constraint would not conform with EAI-1. In summary, enactive AI complements and extends embodied AI's approach to move sense-making into the sensorimotor loop, by grounding sensorimotor interaction in intentional agency (Froese and Ziemke, 2009).

Empowerment as Intrinsic Motivation for Enactive Artificial Agents

We suggest that empowerment maximisation, a principle introduced by Klyubin et al. (2005a), represents a promising candidate for a genuinely intrinsic value function in enactive AI. We will briefly provide the reader with an intuition and formal definition of empowerment and the principle of empowerment maximisation. We will then argue that empowerment supports the formation of constitutive autonomy in enactive agents in both a synthetic and self-constituting manner, and fulfils the requirements for adaptivity without further modifications.

Empowerment and Empowerment Maximisation

Empowerment, the quantity underlying the maximisation principle, is defined over the relationship between an agent's actuators and sensors, and as such is sensitive to the agent's embodiment and Umwelt. It measures the influence of an agent's actions on its environment (controllability), and the extent to which it can perceive this influence afterwards (observability). In other words, empowerment quantifies the options available to an agent in terms of availability and visibility; it measures how much potential influence an agent has on the world it perceives. Klyubin et al. (2008) introduce the principle, while Salge et al. (2014b) provide an extensive survey of motivations, intuitions and past research.

At the centre of the empowerment definition is the interpretation of an agent's embodiment as an information-theoretic communication channel. For any arbitrary separation between an agent and a world we can define sensor variables S and actuator variables A as those states that allow for the in- and outflow of information to the agent, respectively. This interaction with the world is usually described as a perception-action loop (Fuster, 2001; Touchette and Lloyd, 2000, 2004) as in Fig. 1, which can be analysed by means of a causal Bayesian network and Pearl's interventional calculus (Pearl, 2000). Here, arrows imply causation between random variables: the agent's actions A only depend on its sensor input S, which in turn is determined by the rest of the system R. The latter is affected by the preceding system state and the agent's actions. The interventional causal probability distribution $p(S_{t+1}|S_t, A_t)$ thus represents the (potentially noisy) communication channel between actions and future sensor states. For simplicity, the interaction presented here is discrete in time and space. Continuous implementations exist, e.g. for robotics (cf. Salge et al., 2014b).

Empowerment is then defined as the maximum potential information flow (Ay and Polani, 2008) that could possibly be induced by a suitable choice of actions, in a particular state s_t. This can be formalised as the channel's capacity:

$$\begin{aligned}\mathfrak{E}_{s_t} &= \max_{p(a_t)} I(S_{t+1}; A_t) \\ &= \max_{p(a_t)} H(S_{t+1}) - H(S_{t+1}|A_t)) \\ &= \max_{p(a_t)} \sum_{\mathcal{A},\mathcal{S}} p(s_{t+1}|s_t, a_t) p(a_t) \log \frac{p(s_{t+1}|s_t, a_t)}{\sum_{\mathcal{A}} p(s_{t+1}|s_t, \hat{a}_t) p(\hat{a}_t)}\end{aligned}$$

Here, $I(S_{t+1}; A_t)$ represents the mutual information between sensors and actuators, which is based on the difference of regular $H(S_{t+1})$ and conditional Shannon (1948) entropy $H(S_{t+1}|A_t)$. The channel capacity is computed by finding the action distribution that maximises the mutual information. Note that this distribution just defines what the capacity is, and is not the actual action policy. For more information on these notions see (Cover and Thomas, 1991).

Empowerment is *local*, i.e. the agent's knowledge of the local dynamics $p(S_{t+1}|S_t, A_t)$ is sufficient to calculate the quantity. The information-theoretic grounding makes it *domain-independent* and *universal*, i.e. it can be applied to every possible agent-world interaction, as long as this interaction can be modelled as a perception-action loop. This implies that empowerment can be computed on arbitrary agent morphologies, and can cope with changes being made to it. Because the perception action loop can be applied to subsystems (cf. Fuster, 2001), or to formalisations on different levels of abstraction (choosing a more or less fine grained model of what actions and sensors are), empowerment can also be applied to an agent on different hierarchical levels. Finally, empowerment is task-independent, i.e. it is not evaluated in regard to a specific goal or external reward.

Given that empowerment does not measure an agent's actual, but potential influence on the environment, an agent can choose its actions accordingly, in order to get into states with

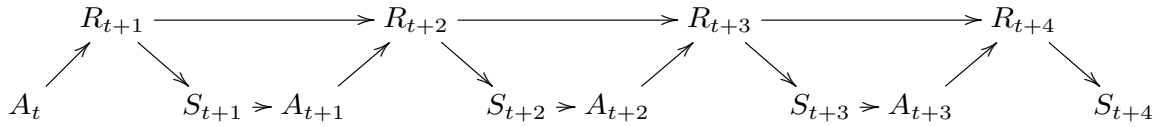

Figure 1: Causal Bayesian network of a memoryless perception-action loop unrolled in time, with the agent's sensors S, actuators A and the rest of the world R.

maximum empowerment. The hypothesis behind this maximisation principle suggests that in order to adapt to changes in their environment, living beings tend to keep their options open. In other words, in absence of more specific goals, they prefer states in which their actions have the strongest potential influence on the environment. More informally, an empowerment-driven agent wants to be in a state where its different actions would have different effects on the world, but it does not necessarily act out all options. This goes hand in hand with a second hypothesis, namely that evolution favoured organisms with efficient information processing (Polani, 2009). Empowerment can thus be understood as one information efficiency principle focusing on the interplay of actuators and sensors. Based on the properties outlined in the previous paragraph, empowerment maximisation satisfies the criteria for an intrinsic motivation function as suggested by Oudeyer and Kaplan (2008).

Empowerment and Constitutive Autonomy

The enactive AI framework suggests that, in order to generate its own identity, a system must continuously maintain its precarious existence (Jonas, 1968; Froese and Ziemke, 2009). Crucially, empowerment maximisation will not, of itself, bring about such a precarious existence. Nevertheless, it allows for its maintenance and supports the process of second order engineering for the emergence of constitutive autonomy. In other words, agents which maintain their empowerment above zero realise *organisational closure*. To support this claim, we will first show that empowerment serves as a proxy for an agent's internal organisation.

A Proxy for Internal Organisation Given that a precarious existence is essentially conditioned on material and energetic requirements, we suggest that an agent's internal processes should maintain its ability to satisfy these requirements. Consequently, an agent has to maintain its capacity to interact with the world by changing and observing it. Empowerment quantifies this capacity; it is non-negative, continuous, and becomes zero if an agent has no influence over the world it perceives. Given that maintaining the ability to interact should be the prime objective, we infer that zero empowerment will inevitably lead to disorganisation. As the internal processes are dependent on these energetic and material requirements, we also deduce that the organisation is impossible to recover without external support. An empowerment value of zero therefore marks the *viability boundary* of an agent and serves as proxy for its internal organisation. It does not capture an agent's precarious existence directly, but the extent to which this existence could be autonomously maintained by means of sensorimotor interaction.

Empowerment does not distinguish whether it is the agent itself which looses coherence or its surrounding world. This is consistent with the theory of situated and embodied cognition, which does not allow separation of the two in terms of their contribution to cognitive processes. In order to maintain its existence, an agent has to keep both its internal processes and its surroundings organised, which is reflected in non-zero empowerment. Crucially, it is guaranteed to become zero if an agent's precarious existence is lost from the point of view of autonomous regeneration, even if its internal organisation is still intact. This provides us with an alternative definition of *death*, which accounts for external forces. For instance, deactivating a robot would result in an empowerment value of zero, because there is nothing the robot could do in order to regain control over its sensorimotor loop, which is in turn required to maintain its existence. Consequently, an empowerment maintaining robot would try to hinder an external force from shutting it down.

If the robot is deactivated nonetheless, the only option then is to rely on an external intervention to bring this capacity back. In a system where the internal organisation relies on a multitude of different active processes, the loss of some causes a chain reaction were others break down, leading to decay of the agent, as it is helplessly exposed to the entropy of the world. Seligman (1975) describes this situation in psychological terms, i.e. from a human perspective. In a classical robot, turning it off is usually not as problematic, as most current robots do not rely on the need to continuously maintain and repair their systems. Nevertheless, if a robot is not turned back on, entropic processes will eventually obliterate the robot, leading to its information-theoretic death, i.e. a complete loss of organisation. In summary, an empowerment of zero marks death in terms of the inability to recover autonomously. This eventually leads to information-theoretic death, which cannot be reversed even by means of external intervention.

Also note that in contrast to other homeostatic variables, such as a robot's energy level, this equation between death and empowerment holds in both ways. A robot could be turned off whilst its energy level, an essential variable for its successful operation, remains high. But a robot cannot be turned off without its empowerment dropping to zero.

Second-Order Homeostasis An empowerment value over the viability boundary thus reflects an agent's efforts to maintain its internal organisation and to sustain its influence on the environment. This is the case even if we assume empowerment to be a meta-variable, i.e. if its maintenance is implemented via several homeostatic processes. Since keeping empowerment non-zero means to keep the agent's internal organisation coherent, which in turn means to keep empowerment non-zero, we end up with a self-referential process. In other words, maintaining empowerment means preserving the capacity to maintain empowerment. It is this form of second order homeostasis that characterises autopoiesis: "an homeostatic (...) system which has its own organisation (...) as fundamental variable which it maintains constant" (Maturana and Varela, 1980, p. 79). Since we are particularly interested in non-physiochemical systems, we make the more general claim that a system which keeps its empowerment non-zero realises *organisational closure*.

Synthetic vs. Second Order Engineering An empowerment maintaining system does not necessarily have to face precarious conditions; the latter must be specified or emerge from the agent's dependencies on the environment. Nevertheless, empowerment can serve as a meta-variable to inform the design of self-constituting agents both from a synthetic and emergent perspective. If we take the earlier, weaker stance of embodied AI and allow for some direct influence on the agent design, empowerment can be used as an explicit intrinsic value function. Maintaining a more specific variable such as an agent's energy level is not sufficient to ensure the coherence of the overall organisation, and thus does not suffice for organisational closure. If we assume empowerment to be implicitly implemented by several other variables, maintaining empowerment in turn means to maintain all variables that are required to keep the organisation coherent. We thus adopt the claim that empowerment "might contribute to modulate pre-imprinted drives or help constituting new homeostatic drives" (Klyubin et al., 2008).

If we stick to strict second order engineering of emergence, empowerment can act as a primer to inform the environmental conditions required for the emergence of a self-constituting agent. More specifically, the environmental conditions must give rise to regulative processes that implement dynamics similar to empowerment maximisation, in order to allow for the emergence of specialised homeostatic variables (Klyubin et al., 2008). As a meta-variable, empowerment allows us to make less explicit assumptions about the specialised processes which must emerge from the environment to constitute and maintain an agent's identity, and yet remains specific enough to enable a more directed process. Counterintuitively, designing the environment in a way that affords the emergence of an empowerment maintaining agent thus allows for more, not less, freedom in emergence and is therefore in sync with the enactive AI principles.

A Sufficiently Intrinsic Value Function Although empowerment is not a truly emergent property in this context, we argue that it is still sufficiently intrinsic to satisfy Froese and Ziemke's requirements for an agent's value function. It is local, and domain-independent through its information-theoretic grounding. This also makes it independent from any sensory semantics, a criterion brought forward by Oudeyer and Kaplan (2008). Embedded in the architecture of a minimal agent with a precarious existence, empowerment becomes grounded in the maintenance of its identity. Calculating empowerment either explicitly or implicitly then translates to assigning genuine relevance to features of the environment.

Empowerment and Adaptivity

We have demonstrated that keeping empowerment non-zero already satisfies a minimal form of adaptivity in terms of maintaining a precarious existence. We will show that this mechanism represents an abstraction of *empowerment maximisation*, a principle which naturally emerges from an agent's need to optimise the efficiency of its sensorimotor interaction. Crucially, empowerment maximisation realises adaptivity without adding additional complexity, e.g. more layers to an agent's architecture.

Distinguishing Viability Tendencies Di Paolo (2005) defines adaptivity as a system's capacity to regulate its states and its relation to the environment with the result that:

1. Tendencies are distinguished and acted upon depending on whether the states will approach or recede from the boundary and, as a consequence,
2. Tendencies of the first kind are moved closer to or transformed into tendencies of the second and so future states are prevented from reaching the boundary with an outward velocity (Di Paolo, 2005).

By quantifying the efficiency of the perception-action loop for different reachable sensor states, empowerment allows the agent to identify states that afford it more options relative to its sensorimotor equipment. Given the link between the agent's internal organisation and the efficiency of its sensorimotor loop, empowerment allows the agent to distinguish tendencies in the environment in terms of how they could potentially affect its viability, which satisfies Di Paolo's first requirement.

We want to stress that the agent does not need to possess a "viability set" in Di Paolo's sense, i.e. different degrees or different forms of disorganisation above its viability boundary. Unlike the value function in embodied AI (A-8), empowerment is future-directed and can therefore differentiate *genuine tendencies* in terms of action affordances that might have an impact on an agent's viability, even if there is no actual robustness in the agent.

As a necessary requirement for real-world scenarios, its information-theoretic foundation enables it to cope with uncertainty in the sensorimotor loop. Anthony et al. (2008) show that empowerment allows an agent to extract and use local information to learn about the world's global structure. An agent which improves empowerment locally in terms of time and space is thus likely to improve globally as well.

Transforming Viability Tendencies Using empowerment maximisation as an action policy allows an agent to prevent states which might prove fatal, and to prefer those which might be beneficial. In a simulation study, we demonstrated that empowerment maximising agents were able to maintain their precarious existence even under serious energy resource constraints (Guckelsberger and Polani, 2014). With empowerment becoming zero when the agent has no sensorimotor control, there is no need to explicitly define a *death state*, and empowerment maximisation naturally leads to death avoidance behaviour. We conclude that empowerment maximisation fulfils Di Paolo's aforementioned, second requirement for adaptivity in terms of *sensorimotor coordination* (Di Paolo, 2005).

Several studies have investigated how empowerment maximisation can facilitate sensorimotor adaptation. Klyubin et al. (2005b, 2008) show that empowerment can serve as an immediate guide for sensor and actuator evolution during an agent's lifetime. They have used empowerment as the fitness function in a genetic algorithm to evolve both sensors and actuators, while constraining the agent's information processing bandwidth. This empowerment maximisation strategy yielded sensors and actuators of different qualities which were "meaningful" in respect to the agent's current state. This is possible because empowerment is not only well defined for different agent morphologies, but even makes these morphologies comparable in terms of which is the better fit for a given environment.

The information-theoretic nature of empowerment allows for a less-biased and thus pro-enactivist approach to sensorimotor adaptation, because it does not rely on any assumptions about sensory modality (Oudeyer and Kaplan, 2008; Salge et al., 2014b). Due to its grounding in the sensorimotor loop, empowerment can potentially be used to modify the environment (Salge et al., 2014a), the agent's morphology, and its sensors and actuators (Klyubin et al., 2008). Hence it also satisfies Di Paolo's requirement for an agent to regulate not only its *states*, but also its *relation* to the environment.

Given the evidence above, we conclude that empowerment maximisation satisfies Di Paolo's requirements for adaptivity. It even exceeds them in that it allows for sensorimotor coordination and adaptation not just in "some circumstances", as Di Paolo (2005) requires, but in a permanent fashion. An empowerment maximising agent not only acts when there is a disaster, but continuously optimises its mastery of the sensorimotor loop. If the empowerment gradient is less steep, empowerment allows for more freedom in the selection of actions.

Discussion

We have claimed that an empowerment maintaining agent can be considered as implementing organisational closure. Nevertheless, we have not yet demonstrated that it meets Maturana's and Varela's second requirement for autopoiesis, namely to constitute itself "as a concrete unity in the space in which the components exist (...)" (Maturana and Varela, 1980, p. 79). Froese and Ziemke point our that there is no mechanism available yet to test for this criterion in non-biophysiological systems (Froese and Ziemke, 2009). Thus, our argument so far is based on the assumption that such a boundary has been somehow established; and we demonstrated how empowerment scales, i.e. that it can be applied to an arbitrary chosen boundary since it is defined on any possible morphology. It is unclear though whether this boundary is maintained for an empowerment maximising agent emerging from second order engineering.

Empowerment maximisation overcomes Wheeler's "intra-context frame problem" Wheeler (2008), i.e. a system's challenge to act appropriately and flexibly in a given context, by assigning potential future states relevance relative to its identity. Nevertheless, in order to maximise empowerment, an agent must infer not only potential future sensor states, given the current state, but also its action consequences in these possibly remote states. The obvious question arising from this is whether computing empowerment, or more broadly speaking, behaving as if one was maximising the empowerment, would require an explicit forward model. Most existing work assumes a somewhat acquired world model that can be queried (Salge et al., 2014b) but more recent work argues that a neural network can be trained to act as if it was maximising empowerment, without an explicit forward model, based only on past experience (Mohamed and Rezende, 2015). In any case it should also be noted that the formalism only requires an agent-centric understanding of the local dynamics $p(S_{t+1}|S_t, A_t)$ based on a level of "representation" consistent with the idea of sensorimotor contingencies (O'Regan and Noë, 2001), i.e. an understanding of the regularities of the agent's own sensorimotor loop.

Revisiting enactive AI's design principles through the lens of empowerment yields that they cannot be as clearly separated as Froese and Ziemke suggest; there must be an implicit value function already in place to maintain the constitutive autonomy of an agent. Adaptivity could resort to the same value function, if the latter is powerful enough to distinguish different viability tendencies. This is the case for empowerment, which scales seamlessly across both requirements without further modifications.

Our investigations also shed light on the issue of robustness: while Froese and Ziemke take physical robustness, i.e.

the existence of a set of non-fatal events, for granted in autopoietic systems, we believe that in the realm of artificial agents, we must allow for systems which can disintegrate in an instant. One might argue that this does not allow for an Umwelt to be constituted, but this is only correct if we think of a value function in embodied AI's terms, determining "whether an action *was* good or bad". Empowerment as a future-directed motivational function in turn allows an agent to distinguish genuine *tendencies* of states to impact its organisation in a positive or negative way. A stochastic system allows for the emergence of such tendencies without the need for the agent to have actual physical robustness. For instance, consider an agent moving across a narrow bridge under windy conditions. Even if the agent only had a binary viability set, it could consider a position at the bridge's edge as more risky, since the likelihood of being blown away is higher, which would eventually render the agent unable to act. Given that such tendencies allow agents to assign relevance to features of the world, i.e. to construct an Umwelt, we suggest that adaptivity is not absolutely necessary for sense-making. Nevertheless, we agree that it is extremely useful in order for agents to improve and compensate.

Conclusion

We have demonstrated that empowerment satisfies the requirements for enactive AI, i.e. constitutive autonomy and adaptivity. We approached these requirements separately, and suggested empowerment as an implicit or explicit, but genuinely intrinsic value function which overcomes the limitations of embodied AI. In particular, we argued that sustaining empowerment is a self-referential process, and that empowerment-driven agents are thus autopoietic.

We demonstrated that empowerment maximisation cannot afford a precarious existence itself, but represents a generic mechanism which ensures the maintenance of such an existence. We believe that empowerment can be realised by means of more specialised variables, or lead to the formation of such variables. By describing how empowerment could support the process of second order engineering for the emergence of constitutive autonomy, we also want to stress its potential role as a mediator between the synthetic methodology of embodied AI and the strict ideas of emergence in enactive AI.

When embedded into an agent with a precarious existence, empowerment will be grounded in the maintenance of its identity. If we take Froese and Ziemke's claims seriously, we can thus assume that the relevance which empowerment assigns to states of the world represents genuine concern. We showed that the principle of maintaining empowerment, as required for constitutive autonomy, is simply a special case of maximising it. Additional layers in an agent's architecture therefore become obsolete: empowerment maximisation represents a mechanism which satisfies the conditions for adaptivity and thus allows an agent to regulate it states and its relation to the environment to move away from its viability boundary.

Froese and Ziemke developed the framework of enactive AI to advance intentional agency and sense-making in artificial agents, and suggest that their requirements represent necessary, although potentially not sufficient conditions. We argue that the second requirement of adaptivity is actually not necessary for sense-making, but extremely useful for the constitution of advanced behaviour and a robust identity. Although they want to move away from carbon chauvinism and Dreyfus' requirement to reproduce living agents in detail (cf. Dreyfus, 2007), their examples in (Froese and Ziemke, 2009) are largely simulations of biochemical processes. We believe that minimal agents motivated by an appropriate intrinsic motivation, such as empowerment, can serve as an inspiring abstraction, which could particularly support the selection of environmental conditions in second order engineering of emergence.

Acknowledgements

CG is funded by EPSRC grant [EP/L015846/1] (IGGI) and CS is funded by the H2020-641321 socSMCs FET Proactive project. The authors would like to thank Martin Biehl, Janet Gibbs, David Lagnado and Daniel Polani for their useful comments and feedback.

References

Anthony, T., Polani, D., and Nehaniv, C. L. (2008). On Preferred States of Agents: How Global Structure is Reflected in Local Structure. In *Proc. 11th Int. Conf. Simulation and Synthesis of Living Systems*, pages 25–32. MIT Press.

Ay, N. and Polani, D. (2008). Information Flows in Causal Networks. *Advances in Complex Systems*, 11(1):17–41.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience, 99th edition.

Di Paolo, E. A. (2005). Autopoiesis, Adaptivity, Teleology, Agency. *Phenomenology and the Cognitive Sciences*, 4(4):429–452.

Dreyfus, H. L. (2007). Why Heideggerian AI Failed and How Fixing it Would Require Making it More Heideggerian. *Philosophical Psychology*, 20(2):247–268.

Froese, T. and Ziemke, T. (2009). Enactive Artificial Intelligence: Investigating the Systemic Organization of Life and Mind. *Artificial Intelligence*, 173(3-4):466–500.

Fuster, J. M. (2001). The prefrontal cortexan update: Time is of the essence. *Neuron*, 30(2):319–333.

Gallagher, S. (2005). *How the Body Shapes the Mind*. Oxford University Press, Oxford University Press.

Guckelsberger, C. and Polani, D. (2014). Effects of Anticipation in Individually Motivated Behaviour on Survival and Control in a Multi-Agent Scenario with Resource Constraints. *Entropy*, 16(6):3357–3378.

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J., and Davidson, R. J. (2010). Cosmetic Use of Botulinum Toxin-A Affects Processing of Emotional Language. *Psychological Science*, 21(7):895–900.

Jonas, H. (1968). Biological Foundations of Individuality. *International Philosophical Quarterly*, 8(2):231–251.

Jonas, H. (1982). *The phenomenon of Life*. University of Chicago Press.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005a). All Else Being Equal Be Empowered. *Advances in Artificial Life*, 3630:744–753.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005b). Empowerment: A Universal Agent-Centric Measure of Control. In *Evolutionary Computation*, pages 128–135.

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):1–14.

Maturana, H. R. and Varela, F. J. (1980). *Autopoesis and Cognition*. D. Reidel.

Maturana, H. R. and Varela, F. J. V. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Shambhala Publications.

Mohamed, S. and Rezende, D. (2015). Stochastic Variational Information Maximisation. *Proc. 29th Conf. Neural Information Processing*, pages 1–9.

O'Regan, J. K. and Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(05):939–973.

Oudeyer, P.-Y. and Kaplan, F. (2008). How Can We Define Intrinsic Motivation? In *Proc 8th Int. Conf. Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 93–101.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pfeifer, R., Iida, F., and Bongard, J. (2005). New Robotics: Design Principles for Intelligent Systems. *Artificial Life*, 11(1-2):99–120.

Pfeifer, R. and Scheier, C. (2001). *Understanding Intelligence*. MIT Press.

Polani, D. (2009). Information: Currency Of Life? *HFSP*, 3(5):307–316.

Rosch, E., Thompson, E., and Varela, F. J. (1992). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Salge, C., Glackin, C., and Polani, D. (2014a). Changing the environment based on empowerment as intrinsic motivation. *Entropy*, 16(5):2789.

Salge, C., Glackin, C., and Polani, D. (2014b). Empowerment – an Introduction. *Guided Self-Organization: Inception*, pages 67–114.

Seligman, M. E. P. (1975). *Helplessness: On Depression, Development, and Death*. WH Freeman.

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

Stewart, J. (2010). Foundational Issues in Enaction as a Paradigm for Cognitive Science: From the Origin of Life to Consciousness and Writing. In Stewart, J., Gapenne, O., and Di Paolo, E. A., editors, *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and Facilitating Conditions of the Human Smile: a Nonobtrusive Test of The Facial Feedback Hypothesis. *Journal of Personality and Social Psychology*, 54(5):768–777.

Touchette, H. and Lloyd, S. (2000). Information-Theoretic Limits of Control. *Physical Review Letters*, 84(6):1156–1159.

Touchette, H. and Lloyd, S. (2004). Information-Theoretic Approach to the Study of Control Systems. *Physica A: Statistical Mechanics and its Applications*, 331.

Von Uexküll, J. (1982). The Theory of Meaning. *Semiotica*, 42(1):25–78.

Wheeler, M. (2005). *Reconstructing the Cognitive World: The Next Step*. MIT Press.

Wheeler, M. (2008). Cognition in Context: Phenomenology, Situated Robotics and the Frame Problem. *Int. Journal Philosophical Studies*, 16(3):323–349.

Zahedi, K. and Ay, N. (2013). Quantifying Morphological Computation. *Entropy*, 15(5):1887–1915.