Addressing the "Why?" in Computational Creativity: A Non-Anthropocentric, Minimal Model of Intentional Creative Agency

Christian Guckelsberger[1], **Christoph Salge**[2,4] and **Simon Colton**[1,3]
[1]Computational Creativity Group, Goldsmiths, University of London, UK
[2]Adaptive Systems Research Group, University of Hertfordshire, UK
[3]The Metamakers Institute, Falmouth University, UK
[4]Game Innovation Lab, New York University, USA
c.guckelsberger@gold.ac.uk

Abstract

Generally, computational creativity (CC) systems cannot explain *why* they are being creative, without ultimately referring back to the values and goals of their designer. Answering the *"why?"* would allow for the attribution of *intentional agency*, and likely lead to a stronger perception of creativity. Enactive artificial intelligence, a framework inspired by autopoietic enactive cognitive science, equips us with the necessary conditions for a value function to reflect a system's own intrinsic goals. We translate the framework's general claims to CC and ground a system's creative activity intrinsically in the maintenance of its identity. We relate to candidate computational principles to realise enactive artificial agents, thus laying the foundations for a minimal, non-anthropocentric model of intentional creative agency. We discuss first implications for the design and evaluation of CC, and address why human-level intentional creative agency is so hard to achieve. We ultimately propose a new research direction in CC, where intentional creative agency is addressed bottom up.

Introduction

Imagine conducting an interrogation experiment, in which human participants are to judge the creativity of a state of the art computational creativity (CC) system. The system could be a piece of software or consist of one or several embodied agents, it could act in the lab or in the field, and there is no restriction on the type of creativity exercised. Crucially, the system has unlimited capacities to enter into a dialogue and to frame (Charnley, Pease, and Colton, 2012) its actions. Participants include the general public, CC researchers, as well as expert practitioners and critics of the type of creativity exercised. In contrast to the Turing (1950) test, the system must always answer truthfully.

We would expect most participants to base their judgement on the system's observed behaviour and produced artefacts only. Some might make few inquiries about the system's process, while others might engage in a deep interrogation. We would certainly end up with divided opinions on the creativity of the system, confirming the view that creativity is an *essentially contested concept* (Gallie, 1955; Jordanous and Keller, 2016). While we would expect most participants to attribute creativity to the system if its behaviour and output was *novel* and *valuable*, others might be more inquisitive, and eventually fail the system because it cannot give satisfactory answers to *why it acted the way it did*.

This addresses the system's *intentional agency*, i.e. its capacity to have a purpose, goal or directive for creative action (cf. Ventura, 2016). However, we doubt that any existing CC system, even with our hypothetical dialogue capacity, could answer questions about its intentionality without referring to its designer's goals. Jordanous and Keller (2016) have empirically identified intentionality as one factor in the perception of creative systems. We believe that a system's inability to account for its *own* intentionality is a valid reason for people to disapprove it of being creative, particularly creative *in its own right*. We also doubt that these systems fully own their artefacts, as they cannot justify why they *originated* them. Ada Lovelace famously addressed originality:

> "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform". (Menabrea and Lovelace, 1842)

By stating that the Analytical Engine has no *pretensions* to originate anything, Lovelace gives us the key to what we believe is the answer to the *"why?" in CC*: if we want to design systems that are deemed creative in their own right, we need them to *own their goals*. Their pretensions, i.e. their motivations, must not be the designers', but arise from their own, genuine *concern*. This concern forms the basis of a system's *sense-making*, i.e. the assignment of values to features of the world that are of relevance to the system itself. To be considered an intentional *agent* in its own right, it must use these values as the basis of action.

Not only do existing implementations fail to address this ultimate challenge of intentional creative agency – CC also misses a theoretical framework describing the conditions for intrinsic goal-ownership underlying intentionality. We believe that the development of such a framework is hindered by CC's focus on human creativity in system design and evaluation. Human creativity unfolds within a complex network of influences shaped by a person's social and cultural environment (Bown, 2015; Jordanous, 2015). Identifying why a person was being creative and translating the findings to formal models therefore is hard. We also believe that CC's focus on big-C artefacts (Kaufman and Beghetto, 2009) is detrimental, as the values within are hard to disentangle and invite complex interpretations of the notion of creativity. Despite these impediments, CC's major contributions to key concepts around intentionality such as *adaptivity*

(Bown, 2012) and *agency* (Bown, 2015) are still strictly anthropocentric. The concept of creativity is human-made, but it should not remain human-centric: by understanding how intentional creative agency can be brought about in artificial agents, we can identify creativity in systems that previously remained unnoticed, learn about new forms of creative behaviour, and actively support their emergence.

We adopt Froese and Ziemke's (2009) *enactive artificial intelligence* (AI) framework, which provides a non-anthropocentric account of intentional agency. In contrast to Dennett's (1989) *intentional stance* which could be applied to any system, Froese and Ziemke (2009) limit intentionality to systems that share some essential characteristics with living organisms. Based on the bio-systemic foundations of autopoietic enactive cognitive science, they argue that the *purpose* of an intentional agent, determining its intrinsic goals, is the maintenance of its existence. They propose two conditions for intentional agency and sense-making in artificial agents: *constitutive autonomy* and *adaptivity*.

We argue that adaptive and constitutively autonomous agents must necessarily exhibit behaviour which many would deem creative. More specifically, we claim that two forms of creativity, *autopoietic-* and *adaptive creativity*, are intrinsic to enactive artificial agents. We hypothesise how our minimal model could give rise to more complex forms of creative behaviour, and briefly outline computational principles to put our theoretical considerations into practice. We thus extend Froese and Ziemke's (2009) framework with an account for creativity to establish a non-anthropocentric, minimal model of intentional creative agency. Our findings suggest that creativity can be found in any living being, not only in humans and highly developed animals. We discuss first implications of our enactive account for the perception of creativity in nature, for CC evaluation, and for building artificial agents with human-level creativity.

Our model is non-anthropocentric but agent-centric, looking at the value of actions and artefacts from the perspective of the creative agent, in contrast to an external observer. It is minimal in that we account for intentional agency in p-creative behaviour (Boden, 2003) at the *edge of being*, in contrast to big-C, h-creativity shaped in a social context (Saunders and Bown, 2015). Accepting creativity as essentially contested, we encourage notions of creativity without or with externally attributed values; by identifying what is required for intentional creative agency, we do not constrain, but widen the scope of what should be considered creative.

Our ultimate goal is to propose a *new direction for research* in CC, in which intentional creative agency is addressed from the bottom up. We motivate our approach by asking the "*why?*" for existing CC systems with a focus on big-C creativity. However, we do not address how this question could be answered in terms of communication and framing (Charnley, Pease, and Colton, 2012). While we put forward a hypothesis about climbing up the *creativity ladder*, closing the explanatory gap between our model and human-level, big-C creativity is subject to further research and experimentation. We agree with Froese and Ziemke (2009) that their conditions are necessary for intentional agency, but also share their caution that they might not be sufficient.

Climbing Down the Creativity Ladder

Traditionally, much research in CC is software-based and models complex human-level creativity in artistic domains. We show by means of case-studies and by reference to traditional AI arguments that this symbolic approach cannot possibly account for intrinsic goal-ownership. Embracing the paradigm of embodied and situated cognition reduces this challenge, but is still insufficient. Our solution eventually leads us away from human, big-C artefacts down to creative behaviour in minimal agents with a precarious existence.

Symbolic Computational Creativity

Developed for more than a decade, *The Painting Fool (TPF)* is a prime example of a symbolic CC system. The project's goal is to create a system which is eventually taken seriously as a creative artist in its own right (Colton, 2012). *TPF* was used extensively in field studies about the perception of creativity by unbiased observers. In the "You Can't Know my Mind" exhibition (Colton and Ventura, 2014), the software accompanied its portraits with a commentary on their creation, enabling visitors to project intentionality on the system. We use this context for a case-study in which *TPF* is equipped with our hypothetical dialogue system.

Being asked "*Why* did you paint my portrait in that way?", *TPF* could genuinely answer: "Because I was reading the following newspaper article, and this was used to simulate a mood which drove the artistic choices I made". Digging deeper, the next question could be: "*Why* were you reading the newspaper?". The answer to this would be: "Because my programmer told to me to do so". A particularly curious participant might then ask: "*Why* did your programmer tell you to read the newspaper?", to which the system would respond: "So that I have an interesting backstory for my creative acts". Asked "*Why* do you need such a backstory?", the system's honest answer would be "So that I appear more creative". We see that our persistent questioning yields a circularity with the reason behind certain behaviours being to promote the appearance of creativity. It is fair to say that *TPF* was given the ability to sustain the impression of intentionality at only *the first level*.

Cook and Colton (2015) account for successful answers one level below. To overcome randomness and hard-coding, they introduce a method to invent distinct, but consistent and therefore believable preferences. Being asked why it painted a portrait in a certain tone, *TPF* could truthfully explain its behaviour through a set of initially generated colour preferences residing within the system. However, asking why it came up with a certain set of preferences, or with the constraints for their consistency, we would again end up in circularity. We believe that many would consider this circularity an unsatisfactory answer to the *"why?"* in CC, and potentially not attribute creativity to such a system.

By asking the *"why?"* for the prototype algorithms described by Ventura (2016), we find that this shortcoming applies to symbolic CC in general. In CC, being intentional is understood as having a goal or directive for action. These goals are modelled with value functions, used in action selection and the post-hoc evaluation of artefacts. There is agreement that in human creativity, such values do not come from the creative person alone, but from a network of human influences (cf. Bown, 2015). In symbolic CC however, a system's goals come exclusively from *other human actors*, and do not reside *in the system itself*: *TPF*'s goals are determined by the algorithms and constraints specified by its designer, and its simulated mood depends on the author of the newspaper article it analyses. Value functions in symbolic CC do not reflect a system's own goals. Moreover, they typically reflect the designer's goals in *respect to a particular artefact*. The system's purpose is determined by the purpose of the artefacts produced (cf. Gervás and León, 2016). The concept of intentionality in symbolic CC is thus very weak.

One might argue that CC software simply does not go "deep enough", but its symbolic nature makes it fundamentally incapable of intrinsic goal-ownership. These systems are *computationalist*, in that creativity is reduced to the manipulation of symbols. Computationalism is subject to a range of classic AI problems such as the *frame problem* (Wheeler, 2005, p. 179) and the *symbol grounding problem* (Harnad, 1990). Searle (1980) addresses the latter in his famous *Chinese Room argument*, showing that syntax is not sufficient for semantics. By translating Searle's argument to artefact creation, Al-Rifaie and Bishop (2015) show that it also applies to CC. Because symbolic CC cannot ground meaning, it also cannot give rise to goals which are meaningful from the system's own perspective.

Embodied Computational Creativity

Brooks (1991) has challenged these problems of symbolic AI by embracing the ideas of situated and embodied cognition. In situated cognition, cognitive processes emerge from the interaction of an organism and its world, and are thus inseparable from action. Embodied cognition emphasises the role of an agent's physical body in shaping cognitive processes. Potentially influenced by *systems theories of creativity* (cf. Saunders, 2012), CC has adopted the embodied and situated approach: there is general agreement that creativity does not occur in a vacuum: it is a *situated* activity, in that it relates to a cultural, social and personal context. However, it is also physically conditioned on an agent's *embodiment* and structured by how an agent's morphology, sensors and actuators shape its interaction with the world.

Embodied AI has developed into a mature framework for modelling artificial agents, and Pfeifer, Iida, and Bongard (2005) describe its characteristics via a list of design principles. Most importantly, they require an agent to have a value function, telling it whether an action was good or bad. The agent must then use these values to motivate its *behaviour*. Embodied AI's value principle thus operates one level below the value function in symbolic CC, which is primarily used in *artefact evaluation* to assess the success of generative routines. We can see the embodied AI principles being adopted in CC: Hoffman and Weinberg (2010) for instance leverage the effect of an agent's morphology on creativity. Their robot Marimba player *Shimon* improvises in real-time to a human pianist's performance. In contrast to symbolic CC, music here is not understood as a sequence of notes, but as a choreography of movements constrained by the robot's morphology. Embodied AI counteracts the Lovelace objection in that it demands a reduction of the designer's influence to foster emergent behaviour. Saunders et al. (2010) investigate the emergence of autonomous and creative behaviour from the interaction of agents in *Curious Whispers*, where a society of simple robots generate and listen to tunes.

Embodied AI practitioners such as Dreyfus (1992) claim that the symbol grounding problem can be overcome by embedding agents in a closed sensorimotor loop: an agent perceives the effects of its actuators on the external environment which, via its internal controller, lead to the next action. An embodied agent can avoid the use of internal symbolic representations, by using "the world as its own model" (Brooks, 1991). However, Froese and Ziemke (2009) argue that this only solves the first part of the *frame problem*:

> "Given a dynamically changing world, how is a non-magical system (...) to take account of those state changes (...) 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, p. 179, emphasis added)

They argue that being embedded in a closed sensorimotor loop is not sufficient for an agent to evaluate features of the world relative to its *own* purpose. Embodied AI's value principle is at the centre of their criticism, as it does not preclude the external assignment of values. *Shimon*'s goals for instance are hard-coded, allowing the system to perform a prescribed set of interactions to support the human musician. Once its counterpart deviates from the protocol, the system fails to operate. Shimon does not act for its own purpose.

Intrinsic Motivation to the Rescue?

We can say that *Shimon* is *extrinsically motivated*, as its goals, defining its behaviour, are imposed by its designers. In contrast, agents can also be intrinsically motivated, performing "an activity for its inherent satisfaction rather than for some separable consequence" (Ryan and Deci, 2000). This psychological definition has been complemented with computational approaches, surveyed by Oudeyer and Kaplan (2008). Per definition, these approaches have to rely on agent-internal experience alone, based on the semantic-free relationship between sensors and actuators. Existing models capture drives like *learning progress* or *curiosity*. *Curious Whispers* uses a model of curiosity to make robots seek interesting tunes. Here, interestingness is quantified by mapping a new tune's novelty, relative to past experience, on a Wundt curve (Saunders et al., 2010). The robots thus listen to tunes which are neither too similar, nor too different to those previously experienced. If the tunes of other agents are not interesting enough, the robots create their own.

Formal models of intrinsic motivation ground behaviour in an agent's sensorimotor loop, and thus partly overcome the criticism of embodied AI's value function. Many models of intrinsic motivation are bio-inspired, but can we claim that any intrinsic value function and the emerging behaviour relates to an agent's own purpose? Why would a certain agent be curious and seek for new stimuli, instead of hiding in a dark room? When does a particular motivational model reflect the system's intrinsic goals, not the designer's? We could simply assume agents to share our goals and behave like us. However, if we accept that cognition and thus behaviour is shaped by our embodiment and situatedness, there is no justification for this anthropocentric stance. The previously outlined theories are not sufficient to decide whether an action policy, which formally qualifies as an intrinsic motivation, reflects the intrinsic goals of a given agent.

Enactive Computational Creativity

Embodied and situated cognition overcomes some shortcomings of symbolic CC, but does not account for an agent's own purpose, i.e. *intrinsic teleology*. Despite this, we argue that intentional creative agency is not an infinite regress problem – we can resolve the circularity demonstrated in symbolic and embodied CC by grounding an agent's actions in a model of intrinsic motivation, based on a *suitable* intrinsic value function. In their enactive AI framework, Froese and Ziemke (2009) introduce the missing conditions for such a value function to relate to an agent's own goals. However, their framework focusses on cognition in general. We argue that enactive agents necessarily have to exhibit two specific forms of creativity which can potentially give rise to complex creative behaviour, and thus establish an account of intentional creative agency. We refer to candidate principles to realise these theoretical conditions, hence laying the foundations for a model of *enactive computational creativity*[1].

Enactive Artificial Intelligence

Froese and Ziemke identify the necessary requirements for an intrinsic motivation to reflect an agent's own goals, by looking at how living beings are different from the non-living with respect to their *purpose*. As mainstream biology offers no distinction in respect to purpose, they draw on the biosystemic foundations of *enactive cognitive science*.

Enactivism is a non-reductive, non-representationalist theory of cognition which adopts the embodied and situated paradigm, but additionally grounds cognition in practical activity. Following O'Regan and Noë's (2001) theory of sensorimotor contingencies, Noë (2004) stresses that what we perceive is determined by what we do. At the core of enactivism thus is the idea that individuals do not passively create internal representations of a pre-given external world (Stewart, 2010); through their interaction with the environment, agents enact, i.e. actively construct, their own world of significance (Varela, Thompson, and Rosch, 1991). Enactivism thus roots sense-making in action. Froese and Ziemke follow the autopoietic branch of enactivism which formulates a strong life-mind continuity, and explicitly addresses intentional agency and sense-making (cf. Thompson, 2004).

At the core of autopoietic enactivism are three theories which root intentional agency in the living, and thus shed light on our missing conditions. Kant argues that living systems have a "natural purpose" in that they are "both cause and effect in themselves" (Kant, 1995, §64). He suggests that the intrinsic goal-directedness of living beings arises from their purpose to self-produce (cf. Weber and Varela, 2002). The biologist von Uexküll translates these findings to sense-making by arguing that living beings, through their sensorimotor activity, construct their own, unique perspective on the world based on the requirements of their self-production (Di Paolo, 2003). In what he refers to as *Umwelt*, features of the environment are captured and assigned significance in terms of how they affect the individuals' self-organisation and ongoing preservation. The bio-philosopher Jonas finally provides us with an account of *identity* (Jonas, 1982, p.126). He argues that simple matter and most artificial systems exist without the need or capacity to act. Living beings in contrast have a *precarious existence*, in that they could at any time become a non-being; they have to continuously interact with their environment in order to satisfy their material and energetic requirements. Based on these theoretical underpinnings, Froese and Ziemke (2009) claim that an agent's value function must reflect the concern about its maintenance of identity. Similarly Haselager (2007) argues that "only systems with a self-generated identity can be said to genuinely own and enact their own goals".

The notion of self-production is operationalised by Maturana and Varela's (1987) concept of *autopoiesis*. An autopoietic system represents a minimal living organisation which realises constitutive autonomy, as it physically individuates an entity from its environment, and constitutes its identity in the domain. The concept of autopoiesis only applies to biochemical systems, and is generalised by the notion of *organisational closure*. A system implementing organisational closure is a network of processes that generate and sustain its identity under precarious conditions, and that form a unity in a containing domain (Varela, 1979). The first condition for enactive artificial agents states that intrinsic teleology requires 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 condition represents the enactive version of embodied AI's value principle, but it is strictly intrinsic in that the agent must relate value to the maintenance of its own precarious existence. However, Froese and Ziemke (2009) argue that this alone is not sufficient to maintain an agent's identity over the long term: in a dynamic and uncertain environment, an agent must be able to compensate for unexpected events. This requires a value function to distinguish external events more gradually relative to the agent's organisation. Nevertheless, the concept of organisational closure per se is only binary: a system either maintains its organisational closure or not. Di Paolo (2005) compensates for this limitation by presupposing the existence of a *viability* set, i.e., levels of structural change that allow living beings to "sustain a certain range of perturbations [...] before they lose their au-

[1] With "computational", we relate to CC as a research field, not to computationalism. Our model uses insights from autopoietic enactivism to guide the design and evaluation of artificial, intentionally creative agents. We believe that this approach is not restricted to simulated embodied systems and robotics, but could also inform other means of creating artificial agents, e.g. synthetic biology.

topoiesis" (Di Paolo, 2005). He defines *adaptivity* as

> "a system's capacity, in some circumstances, to regulate its states and its relation to the environment with the result that, if the states are sufficiently close to the boundary of viability,
>
> 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)

The necessity of adaptivity for intentional agency is covered by the second condition for enactive agents:

> **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).

In summary, enactive AI complements and extends embodied AI's approach to move sense-making into the sensorimotor loop, by grounding sensorimotor interaction in an agent's maintenance of its identity. Froese and Ziemke (2009) claim that an intentional agent must not only be embodied, but also realise constitutive autonomy and adaptivity via its value function and action policy.

Creativity at the Edge of Being and Beyond

Two factors can be frequently observed in the attribution of creativity to other humans: novelty and value (Paul and Kaufmann, 2014; Jordanous and Keller, 2016). Current CC systems lack intrinsic goal-ownership, because the discipline misses an account for value from a non-human perspective. When researchers talk about creativity outside the human domain, they simply drop value from the definition: What Boden (2015) labels "biological creativity" in the context of artificial life is earlier defined by Bown (2012) as "generative creativity": a system's ability to create "new patterns or behaviours regardless of the benefit to that system".

Autopoietic enactivism provides us with an intrinsic account of value in any constitutively autonomous system. Although autopoiesis literally means "self-creation", we argue that a system with organisational closure alone cannot be assumed to exhibit *novel* behaviour in a non-trivial sense. We thus distinguish *autopoietic creativity* in organisationally closed systems from *adaptive creativity* in fully enactive agents. We claim that an enactive AI must necessarily exhibit adaptive creativity. The notion "adaptive creativity" has been used loosely in creativity studies (Kirton, 1994) and CC (Bown, 2012) before; we make it more concrete by drawing on Di Paolo's (2005) definition of adaptivity. While the notion of value in human-creativity is ambiguous and relates to complex concepts such as interestingness and aesthetics, enactive AI allows us to root it in utility alone.

Autopoietic Creativity Autopoiesis is generalised in operational closure. Varela (1984) uses the notion of a "creative circle" to describe both (i) the self-organisation of an organisationally closed system, and (ii) its self-maintenance as an autonomous unity. We can distinguish two notions of creativity along these stages. While we witness an autonomous system emerging out of something else, e.g. a cell out of a molecular soup, we can only apply the notion of *generative creativity*. From the perspective of an external observer, the system appears as a transient, ever-changing artefact. At this stage, there is no perspective of the system itself yet, and value can only be externally imposed.

However, once the system has individuated itself from the containing domain, it establishes a unique perspective on the world, mediated by its situatedness and embodiment. We can now consider creativity from the system's *own* perspective. From here, every change to its structure has a value, in that it either preserves or destroys its organisation. To maintain its identity, a system has to engage in positively valued, organisation-preserving operations (Jonas, 1982, p. 72). We define *autopoietic creativity* as a system's *active* modification of stucture to ensure its continuous existence. A system cannot be referred to as autopoietically creative, if these changes are caused exclusively by external forces.

We argue that this form of creativity is very minimal, in that the exhibited behaviour is not necessarily novel in a non-trivial sense. To distinguish autopoiesis clearly from adaptivity, we constrain our claim to a fictional autopoietic system operating free from perturbations, i.e. there are no external forces which could destroy its organisation. For a system to maintain organisational closure, it is sufficient to engage in a cyclic flow of material configurations. In the absence of perturbations, novelty in the system's change of structure only depends on its current shape and its internal dynamics. Consequently, the range of possible changes and thus novelty could be fully pre-specified and would be quickly exhausted. This does not mean that an autopoietic system could not exhibit valuable and novel behaviour; but since this is not required by definition, we cannot assume it to be creative in the popular sense described earlier.

Adaptive Creativity The notion of *autopoietic creativity* is rather theoretical, as a physically embodied system will always be subject to entropic forces, either implicitly via material and energetic dependencies on the environment, or explicitly through perturbations leading to its disorganisation. To maintain its identity, such a system has to be adaptive. We argue that adaptivity represents the essential mechanism for creative behaviour in an autopoietic system.

According to Di Paolo (2005), an adaptive system does not disintegrate in a second, but can undergo a variety of structural changes before it looses its existence. These structural changes characterise the system's *viability set*. Di Paolo argues that an adaptive system must be able to recognise whenever it is moving closer to its viability boundary, and either slow this tendency down, or invert it in order to be more robust against future perturbations. In contrast to the earlier isolated system, the future structure of an adaptive system is not only affected by its current shape and internal dynamics, but also by external perturbations. An adaptive system exhibits *novel* behaviour when it is either (i) responding to a familiar perturbation in a different way than before, or when it is (ii) responding to a previously unencountered perturbation. This is non-trivial as in a dynamic environment, an embodied system cannot be hard-wired to anticipate and defend its identity successfully against any possible perturbation; increasing the complexity of its internal dynamics comes with an increase in energetic and material requirements, and thus counteracts viability. Any constitutively autonomous and adaptive agent must be able to respond flexibly to potentially unencountered perturbations in novel but valuable ways. We define this as *adaptive creativity* and conclude that an enactive agent with intentional agency must necessarily be adaptively creative.

Moving Away From the Edge An adaptive system must be able to evaluate each structural change in its viability set relative to its viability boundary. A structural change that would move the system closer to its viability boundary would be valued negative and vice-versa. Adaptive creativity, i.e. responding with novel and organisation-preserving actions to potentially unencountered perturbations, allows a system to move away from the boundary.

However, Di Paolo (2005) only requires an adaptive system to regulate its states “in some circumstances”. While such a system would likely stay close to its viability boundary, acting *consistently* creative would allow it to move away from the edge of being. We hypothesise that such *aspirational creative behaviour* requires, but also gives rise to more complex forms of creativity. To compensate and escape viability tendencies, a system could change its behaviour via sensorimotor coordination, but it could also adapt and augment is morphology, or change its environment. As Gibson has stated: “Why has man changed the shapes and substances of his environment? To change what it affords him. He has made more available what benefits him and less pressing what injures him” (Gibson, 1986, p. 123). Furthermore, we expect sociality to be particularly important for sustaining viability. We hypothesise that aspirational adaptive creativity allows us to climb up the *creativity ladder* again. However, detailed theoretical work and experimentation is subject to future work.

Operational Principles

There are many models of artificial curiosity, and we cannot make a general claim without evaluating them individually based on the enactive AI conditions. We believe that the intrinsic model of curiosity in *Curious Whispers* (Saunders et al., 2010) cannot give rise to intentional creative agency. Only in the presence of a function indicating the agent’s viability relative to its current state can an agent relate its behaviour to the maintenance of its existence. An agent is not constitutively autonomous, if it does not act based on how close it is to losing its autonomy. In *Curious Whispers*, it is unclear how selecting tunes which are neither too familiar nor too novel relates to a robot’s viability.

Recently, several candidate principles were hypothesised to realise the enactive AI conditions. In his *Free Energy Principle*, Friston (2010) argues that in order to not become disorganised, living organisms and artificial agents have to maintain an upper bound on the entropy of their sensory states as the average of surprise. The agent’s free energy, formalised as the difference between its model and the world, constitutes a tractable upper bound on surprise. Allen and Friston (2016) argue in the context of predictive coding that a free energy minimising agent can be considered as enactive, and that it realises constitutive autonomy and adaptivity. The *Free Energy Principle* evolves around action and model optimality, but it is unclear whether it provides values for sense-making. Friston’s principle appears conceptually close to maximising *predictive information* (Ay et al., 2008), i.e. the information an agent’s past states hold about its future. A comparison of the principles is yet to be done.

Guckelsberger and Salge (2016) argue that the information-theoretic principle of *empowerment maximisation* fulfils the enactive AI conditions. Empowerment quantifies the efficiency of an agent’s sensorimotor loop. Given that an agent could not counteract perturbations and satisfy its energetic or material requirements with a dysfunctional loop, they argue that empowerment represents a proxy to the agent’s viability, and that maintaining it realises organisational closure. They show via simulations that an agent which maximises empowerment in its action policy realises adaptivity. Crucially, an empowerment maximising agent not only adapts sporadically, but consistently increases its viability. We thus consider this a promising candidate to investigate whether consistent adaptive creativity leads to more complex creative behaviour.

Implications: The Embodiment Distance

We have proposed a model of intentional creative agency in which value is grounded in a system’s maintenance of its precarious existence. We argue that a system’s sense-making and thus creative behaviour is determined by its embodiment, which is usually very *different from ours*: a physically embodied agent can have a different morphology, a different access to the world through its sensors and actuators, and other energetic and material dependencies. We briefly discuss first implications of this *embodiment distance* for the perception of creativity in nature and artificial systems, as well as for the design of human-like CC.

CC field studies (e.g. Colton and Ventura, 2014) demonstrate that there are many systems that unbiased observers deem creative, although these systems ultimately do not act creatively in respect to their own goals. This is fair, as creativity is an essentially contested concept. Here, we look at the opposite case: we argue that there are many *adaptively creative* systems which perform novel and valuable actions relative to their intrinsic values, but would *not* be deemed creative. These systems root their values and thus behaviour in their embodiment. Their artefacts, i.e. their own structure and marks in the environment, are consequently value-laden relative to their embodiment. Our judgement of human artefacts is sensitive to our human embodiment, as psychological experiments in embodied aesthetics (Johnson, 2008) suggest. When we evaluate the creativity of non-human systems with intentional agency, we are likely to misjudge value in their behaviour or artefacts, or hesitate to attribute any value at all, as our embodiment distance is too large. This means that we are likely to misjudge or even fail to acknowledge the adaptive creativity of some systems, while agents of the same type would value it highly. To judge the adaptive creativity of a system with creative intentional agency, we need to take the perspective of that system, and assess its sense-making and behaviour from there.

The embodiment distance is also relevant for the design of intentional CC agents with human-like creativity. Dreyfus (2007) has argued that human-like cognition in artificial agents requires us to replicate human embodiment. While this is not an issue if we are only interested in realising adaptive creativity in minimal intentional agents, Dreyfus' claim remains critical for reproducing human-like creativity.

Related Work

We have operationalised the previously loose notions of agency (Bown, 2015), adaptivity (Bown, 2012) and autonomy (Saunders, 2012) in CC by drawing on autopoietic enactivism. We seem to be the first to embrace this branch of enactive cognitive science in CC; Davis et al. (2015) develop a model of creative collaboration and co-creation based on Noë's (2004) sensorimotor enactivism; however, the sensorimotor branch focusses on the constitutive role of action in perception, but misses an account of intentional agency.

The concept of autopoiesis has been used in systems theories of creativity. However, it has been employed rather metaphorically (Gornev, 1997) or to describe creativity in society, not in individual agents (Iba, 2010). CC has adopted the concept of autopoiesis: Bishop and Al-Rifaie (2016) implemented an autopoietic model of creativity as a swarm intelligence system, but they specify their value function explicitly, instead of using an intrinsic account of sense-making. Saunders (2012) investigates the role of communication for autonomy in creative agent societies, but does not ground the behaviour of individual agents in the maintenance of their identity. Most importantly, none of the approaches provides an account of intrinsic teleology and relates it to intentional (creative) agency.

Conclusion and Future Work

We have shown via case-studies that existing CC systems, typically with a focus on human creativity, cannot provide a satisfactory answer to *why* they are being creative because they lack intrinsic goal-ownership. We have adopted the enactive AI framework for a non-anthropocentric account of intentional agency in minimal, embodied agents. Creativity, as commonly perceived, seems to be maximally removed from how simple organisms survive and cope within an ever-changing environment. By showing that constitutively autonomous and adaptive agents must necessarily perform intrinsically valuable and novel actions, we have grounded two of the strongest factors in the attribution of creativity in the essence of the living. It follows that enactive AI's conditions for an intrinsic value function to reflect an agent's own goals are also necessary for intentional *creative* agency. Saunders (2012) notes that AI was missing a means to realise constitutive autonomy. We have referred to operational principles which are hypothesised to also realise adaptivity, and thus laid the foundations for a non-anthropocentric, minimal model of intentional creative agency.

The enactive account to sense-making highlights that we can only assess adaptive creativity in artificial and natural agents with intentional creative agency if we switch perspectives: We have to take *their* embodiment into account, *not ours*. However, it is still to be discussed whether this is subject to the *hard problem* of knowing what it is like to be that system (Nagel, 1974), or if it is sufficient to use introspection in artificial agents to learn about how the system makes sense of its environment as basis for its behaviour. We believe that our non-anthropocentric model can advance progress in CC and extend the scope of the field: if our creativity relies on our embodiment, it is necessarily subject to constraints as in other embodied agents. There might be other forms of creativity in nature, resulting from different constraints, which could benefit us. AI allows us to explore these in simulations even beyond the laws of the physical world. Similar to a famous endeavour in *artificial life* (Shanken, 1998), we encourage looking beyond creativity in nature, and investigate *creativity as it could be*.

We suggest this as a point of departure for a new research direction in CC, addressing intentional creative agency from the bottom up. One of the biggest challenges will be to close the explanatory gap between the creative intentional agency in our minimal agents and humans. As a first step, we have hypothesised that *aspirational*, i.e. consistently adaptive agents will give rise to more complex creative behaviour. To evaluate our model in practice, we have to address the engineering challenge of building a physically embodied agent that can counteract its precarious existence. We agree with Saunders (2012) that more complex forms of creativity require the interaction with other agents, so extending our model to social creativity using insights from enactivism is a promising next step. We also want to refine the sense-making granularity of the current model by drawing on biosemiotic enactivism (De Jesus, 2016). We are fascinated by the following question: If embodied systems establish their own world of meaning relative to their embodiment, what do their creative products and processes look like, and how do they differ from ours? We suggest starting this investigation with minimal, intentionally creative agents and climbing up the "creativity ladder".

Acknowledgments

CG is funded by EPSRC grant EP/L015846/1 (IGGI). CS is funded by Marie Sklodowska-Curie grant 705643. SC is funded by EC FP7 grant 621403 and EPSRC Leadership Fellowship EP/J004049.

References

Al-Rifaie, M. M., and Bishop, J. M. 2015. Weak and Strong Computational Creativity. In *Computational Creativity Research: Towards Creative Machines*. Atlantis Press. 37–49.

Allen, M., and Friston, K. J. 2016. From Cognitivism to Autopoiesis: Towards a Computational Framework for the Embodied Mind. *Synthese* 1–24.

Ay, N.; Bertschinger, N.; Der, R.; Güttler, F.; and Olbrich, E. 2008. Predictive Information and Explorative Behavior of Autonomous Robots. *European Physical Journal B* 63(3):329–339.

Bishop, J., and Al-Rifaie, M. 2016. Autopoiesis in Creativity and Art. In *Proc. Conf. MOCO*.

Boden, M. A. 2003. *The Creative Mind: Myths and Mechanisms*. Routledge, 2nd edition.

Boden, M. A. 2015. Creativity and ALife. *ALife* 21(3):354–365.

Bown, O. 2012. Generative and Adaptive Creativity. In *Computers and Creativity*. Springer. 361–381.

Bown, O. 2015. Attributing Creative Agency: Are we doing it right? In *Proc. 6th ICCC*, 17–22.

Brooks, R. A. 1991. Intelligence Without Representation. *Artificial Intelligence* 47(1-3):139–159.

Charnley, J.; Pease, A.; and Colton, S. 2012. On the Notion of Framing in Comp. Creativity. In *Proc. 3rd ICCC*, 77–81.

Colton, S., and Ventura, D. 2014. You Can't Know my Mind: A Festival of Comp. Creativity. In *Proc. 5th ICCC*, 351–354.

Colton, S. 2012. The Painting Fool. Stories from Building an Automated Painter. In *Computers and Creativity*. Springer. 3–38.

Cook, M., and Colton, S. 2015. Generating Code For Expressing Simple Preferences: Moving On From Hardcoding And Randomness. In *Proc. 6th ICCC*, 8–16.

Davis, N.; Hsiao, C.-P.; Popova, Y.; and Magerko, B. 2015. An Enactive Model of Creativity for Computational Collaboration and Co-creation. In *Creativity in the Digital Age*. Springer. 223–243.

De Jesus, P. 2016. From Enactive Phenomenology to Biosemiotic Enactivism. *Adaptive Behavior* 24(2):130–146.

Dennett, D. C. 1989. *The intentional stance*. MIT press.

Di Paolo, E. 2003. Organismically-Inspired Robotics: Homeostatic Adaptation and Teleology Beyond the Closed Sensorimotor Loop. In *Dynamical Systems Approach to Embodiment and Sociality*. Advanced Knowledge Int. 19–42.

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

Dreyfus, H. L. 1992. *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT press.

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

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience* 11(2):127–138.

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

Gallie, W. B. 1955. Essentially Contested Concepts. *Proc. Aristotelian Society* 56:167–198.

Gervás, P., and León, C. 2016. Integrating Purpose and Revision into a Computational Model of Literary Generation. In *Creativity and Universality in Language*. Springer. 105–121.

Gibson, J. J. 1986. The Theory of Affordances. In *The Ecological Approach to Visual Perception*. Routledge. 127–138.

Gornev, G. P. 1997. The Creativity Question in the Perspective of Autopoietic Systems Theory. *Kybernetes* 26(6/7):738–750.

Guckelsberger, C., and Salge, C. 2016. Does Empowerment Maximisation Allow for Enactive Artificial Agents? In *Proc. 15th Int. Conf. ALIFE*, 704–711.

Harnad, S. 1990. The Symbol Grounding Problem. *Physica D* 42(1-3):335–346.

Haselager, W. F. 2007. Robotics, Philosophy and the Problems of Autonomy. *Pragmatics & Cognition* 13(3):515–523.

Hoffman, G., and Weinberg, G. 2010. Gesture-based Human-Robot Jazz Improvisation. In *Proc. 2010 IEEE Int. Conf. Robotics and Automation*, 582–587.

Iba, T. 2010. An Autopoietic Systems Theory for Creativity. *Procedia - Social and Behavioral Sciences* 2(4):6610–6625.

Johnson, M. 2008. *The Meaning of the Body: Aesthetics of Human Understanding*. Univ. of Chicago Press.

Jonas, H. 1982. *The Phenomenon of Life*. Univ. of Chicago Press.

Jordanous, A., and Keller, B. 2016. Modelling Creativity: Identifying Key Components through a Corpus-Based Approach. *PloS one* 11(10).

Jordanous, A. 2015. Four PPPPerspectives on Computational Creativity. In *Proc. AISB Symp. Computational Creativity*, 16–22.

Kant, I. 1995. *Kritik der Urteilskraft (Critique of Judgment)*. Suhrkamp, 1st 1790 edition.

Kaufman, J. C., and Beghetto, R. A. 2009. Beyond Big and Little: The Four C Model of Creativity. *Rev. Gen. Psych.* 13(1):1.

Kirton, M. J. 1994. *Adaptors and Innovators: Styles of Creativity and Problem Solving*. Routledge.

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

Menabrea, L. F., and Lovelace, A. 1842. *Sketch of the Analytical Engine Invented by Charles Babbage*. Richard and John Taylor.

Nagel, T. 1974. What is it like to be a bat? *The Philosophical Review* 83(4):435–450.

Noë, A. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

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 Conf. Epigenetic Robotics*, 93–101.

Paul, E. S., and Kaufmann, S. B. 2014. Introducing The Philosophy of Creativity. In *The Philosophy of Creativity: New Essays*. Oxford Scholarship Online. 3–16.

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

Ryan, R., and Deci, E. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25(1):54–67.

Saunders, R., and Bown, O. 2015. Computational Social Creativity. *ALife* 21(3):366–378.

Saunders, R.; Gemeinboeck, P.; Lombard, A.; Bourke, D.; and Kocaballi, B. 2010. Curious Whispers: An Embodied Artificial Creative System. In *Proc. 1st ICCC*, 100–109.

Saunders, R. 2012. Towards Autonomous Creative Systems: A Computational Approach. *Cog. Comp.* 4(3):216–225.

Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3(3):417–457.

Shanken, E. A. 1998. Life as We Know It and / or Life as It Could Be: Epistemology and the Ontology / Ontogeny of Artificial Life. *Leonardo* 31(5):383–388.

Stewart, J. 2010. Foundational Issues in Enaction as a Paradigm for Cognitive Science. In *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press. 1–32.

Thompson, E. 2004. Life and Mind: From Autopoiesis to Neurophenomenology. A Tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences* 3(4):381–398.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433–460.

Varela, F. J.; Thompson, E.; and Rosch, E. 1991. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Varela, F. J. 1979. *Principles of Biological Autonomy*. Elsevier.

Varela, F. J. 1984. The Creative Circle. Sketches on the Natural History of Circularity. In *The Invented Reality: Contributions to Constructivism*. WW Norton. 309–325.

Ventura, D. 2016. Mere Generation: Essential Barometer or Dated Concept? In *Proc. 7th ICCC*, 17–24.

Weber, A., and Varela, F. J. 2002. Life after Kant: Natural Purposes and the Autopoietic Foundations of Biological Individuality. *Phenomenology and the Cognitive Sciences* 1(2):97–125.

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