Computational Creativity: A Philosophical Approach, and an Approach to Philosophy

Stephen McGregor, Geraint Wiggins and Matthew Purver
School of Electronic Engineering and Computer Science
Queen Mary University of London
s.e.mcgregor@qmul.ac.uk, geraint.wiggins@qmul.ac.uk, m.purver@qmul.ac.uk

Abstract

This paper seeks to situate computational creativity in relation to philosophy and in particular philosophy of mind. The goal is to investigate issues relevant to both how computational creativity can be used to explore philosophical questions and how philosophical positions, whether they are accepted as accurate or not, can be used as a tool for evaluating computational creativity. First, the possibility of symbol manipulating machines acting as creative agents will be examined in terms of its ramifications for historic and contemporary theories of mind. Next a philosophically motivated mechanism for evaluating creative systems will be proposed, based on the idea that an intimation of dualism, with its inherent mental representations, is a thing that typical observers seek when evaluating creativity. Two computational frameworks that might adequately satisfy this evaluative mechanism will then be described, though the implementation of such systems in a creative context is left for future work. Finally, the kind of audience required for the type of evaluation proposed will be briefly discussed.

Introduction

In quotidian interactions, either on a personal or social level, computers are such familiar devices that their operations are taken for granted as having the same kind of relatively universal grounding that humans engaging in interpersonal exchanges of information employ. When computers become either the platform for or the object of philosophical enquiries, though, it becomes necessary to talk about them as information processing systems or as symbol manipulating machines (per Newell and Simon, 1990): in this sense, the operations which computers perform must be seen as transpiring in an abstract space, defined by a system of information grounded somehow relative to an observer. This quality of computation immediately introduces a problematic element of subjectivity to the assessment of a purely informational system's ability to generate meaning, and an ambiguity arises over whether such a system can really autonomously produce output which has been invested with semantic content.

It is due to precisely this key feature of computational systems, their dependence on an observer for operational coherence, that computers have become an element in various philosophical discussions, often in the form of *reductiones ad absurdem*, exercises aimed at problematising both reductionist and internalist accounts of mental phenomena. Putnam (1988) in particular has argued for the computational significance of the internal states of a rock, while Searle (1990) constructed his famous Chinese room argument to demonstrate the absence of intentionality in machines which merely manipulate symbols, a stance subsequently used as a platform for questioning the very basis of cognitive experience. In these examples, computers come out as the foils for arguments about the intractable difficulty of defining or even talking about human consciousness. Rather than treating computers as the theoretical objects of thought experiments, this paper will argue, as Sloman (1978) did several years ago, that computers should be considered essential tools for doing good philosophy, and that in particular the question of whether computers can be autonomously creative is philosophically valid.

This paper's first objective is to place the field of computational creativity within the context of the philosophy of mind, and in particular to consider how the field might be used as a vehicle for empirically exploring the problem of dualism which has been characteristically at the centre of questions regarding the mind and consciousness in modern Western philosophy. To this end, a strong counterargument to the traditional mode of dualism, which argues that the mind and physical matter occupy two mutually irreducible spaces, can be found in considering ways in which symbol manipulating machines might be able to autonomously produce informational artefacts that are new and valuable and that furthermore bear some sort of meaning relevant to the way in which the creative system itself operates. If a computational system can produce new, valuable artefacts in a way that is deemed suitably creative, and yet these systems are themselves reducible to manipulations of symbols grounded in the workings of a physical machine, there seems to be no case to make for the idea that the act of generating new meaning in the world transpires in some intangible mental domain.

The second objective of this paper is to propose a new mechanism for evaluating creative systems, motivated by insights into the way that humans view themselves. Taking the intransigence of the mind/body problem as a starting point, it will be suggested that it is precisely the kind of represen-

tational internal states that dualists have attributed to the immaterial space of the mind that should be sought in the operations of creative agents. While a positive assessment of the creativity of an informational system would clearly negate the premise of a mental space separate from physical reality, it is argued herein that precisely this negation serves as a good basis for using the mere impression of such states in a system as an ersatz device for evaluating the real presence of creative behaviour. To this end, two topical computational frameworks, vector space models and deep belief networks, will be put forward as candidates for future work in various domains of computational creativity, with the view that these approaches to computation have the potential to build conceptual structures which might be considered by some observers as corresponding to the type of mental representations attributed to humans.

Computational Creativity and the Demise of Dualism

Descartes' (1911) theory of a mind/matter divide, and the notion of internal mental representations which in particular have characterised the type of introspective reports of the mental space described by philosophers of this bent, have been at the centre of the development of modern Western philosophy, with subsequent canonical philosophers routinely name checking Descartes. The dualism inherent in the mind/matter world view has, however, fallen so severely into disrepute with latter day theorists of mind that a cognitive scientist recently felt comfortable in asserting that in the field today, "even the word 'Cartesian' is often used as a term of abuse," (Rowlands, 2010, p 12). Indeed, in their immensely public debate over the nature of consciousness, Dennett and Searle (1995) resort to mutual accusations of existential partitioning, with both thinkers avowing their own faithfulness to what they perceive as the fundamental, indeed, explicitly monist type of data on which an analysis of existence should be based and upon which any theory of consciousness must supervene.

So the great feuds of contemporary philosophers have been characterised not by a debate over the extent of the merits and faults of dualism, but rather by quarrelling about the precise way in which this dead idea should be autopsied. Whether from the material perspective of reductionist science or from the subjective vantage point of emergent intentionality, the idea that the mind inhabits some physically irresolvable realm has been rejected. This rejection has done little, however, to mitigate the deep issues which characterise the problem of cognition. Furthermore, where strong dualism has been largely vanquished from the philosophical vanguard, it seems even more clear that blunt behaviourism has been thoroughly rebuffed: the idea that cognition can be discussed in terms of simply observed bodily reactions is considered philosophically infeasible (see Boden, 2006, for an overview). The mind evidently experiences the world not as simulating data, but as an array of semantically loaded entities that interact on various levels and according to various rules. The consequent problem of what constitutes perceptual cognition has been characterised as "the binding problem", by which the mind must somehow perform the trick of corralling multifarious sensory stimuli into a unified experience of reality consisting of discernible, describable things which exist on various levels of abstraction.

With this in mind, certain radical views are open to misinterpretation as harbingers of a Cartesian resurrection. For instance, Chalmers (1996) describes a nuanced functionalism by which an agent is conscious by merit of the processes that it performs on a certain level of abstraction regardless of the physical mechanisms of those processes, and Pattee (2008) posits that language and physics should be viewed as two intertwined but mutually irreducible phenomena. Humans are somehow engaging in the act of meaning, in the sense that Wittgenstein indicated when he wrote that "only the act of meaning can anticipate reality" (Wittgenstein, 1967, p 76): it is the characteristically human ability to see a world full of meaningful things rather than just a world full of data. It is not clear, however, how the binding problem is solved, and how the multifarious world is transformed from material input into expressions which are likewise fundamentally material through a cognitive process which is somehow perceptive and expressive. The human ability to perpetually perform this trick is the subject of the dispute between Searle and Dennett, and is the object of what Chalmers has characterised as the "hard problem" of consciousness. The answers to these questions remain arguably as opaque as they were in Descartes' world.

It would seem that computational creativity should, in principle, be the darling of any effort to empirically vanquish any remnant of dualism: to show that a physically grounded symbol manipulating machine is capable of participating as an agent in meaning-making interactions could only illustrate the fallacy of the supposition that such things occur in some kind of non-material space. Wiggins (2012) has recently argued that creativity is in fact the substrate of consciousness, with the capacity for an agent to imagine the world as being different than it is serving as the basis for cognitive action in an environment. In this scenario, an information theoretical process corresponds to Wittgenstein's act of meaning, with statistical computations of perceptual data emerging as semantically gravid expectations of what will happen in an environment. Creativity itself becomes precisely Wittgenstein's act of producing new meaning, of building new ways of perceiving and anticipating the world on different levels of abstraction. Notwithstanding the resilient arguments from Searle (1990) that purely informational symbol manipulating systems cannot have intentionality at the root of their machinations, it would seem that just the ability for an algorithmic machine to be creative would at least prove that the basis of consciousness can be in the material world of physics.

So the argument here is not that, in being creative, in participating in the act of meaning, computers have some chance of becoming conscious. Even with this caveat, though, there is an inherent causal ambiguity in the stance that computers can be creative: it is not clear that the idea that machines can autonomously generate meaningful output *a priori* is necessarily sound. In fact, short of imposing emergent phenomenological properties on hardware, ac-

ceptance that a computer can be creative implies a *de facto* rejection of dualism on the grounds that the machine cannot imaginably be partly located in some immaterial mental space. A tautology emerges by which a positive result for computational creativity is dependent on precisely the reductionist premise that it will hopefully be used to prove.

Rejection of dualism and of the corollary representations which inhabit a placeless mental space, on the other hand, do not necessarily entail an acceptance of the idea that computers can participate in the same kind of creative meaning-making as humans. Indeed, a notable trend in contemporary cognitive science is a move away from the idea that symbolic approaches to the mind can have anything to do with cognition at all, as characterised by the work of Noë (2004), Rowlands (2010), and Chemero (2009). This unfolding movement in the theory of mind traces its roots back to the enactivism of Varela, Thompson, and Rosch (1991) and to the ecologically situated psychology of Gibson (1979): these traditions seek to embed the thinking organism in a physical environment from which the processes underlying consciousness cannot be isolated. In terms of building creative machines, this bodily, environmental approach seems to indicate something more like robotics than the traditional conception of computational creativity as involving the algorithmic construction and traversal of abstract informational state spaces per Boden (1990).

Hence, if computational creativity is to be used as a tool for talking empirically about philosophical questions, the burden of proof shifts onto demonstrating somehow or another that information processing systems can behave creatively in the first place. If this can be done, then it seems likely that an analysis of the specific types of systems which generate creative output might yield some interesting philosophical insights into the nature of cognition. But as will be illustrated in the next section, the evaluation of computational creativity is not by any means a straightforward issue.

Evaluating Symbol Manipulating Systems

The problem of evaluation is a significant aspect of Boden's (1990) classic treatment of computational creativity, where it is argued that in order for computer generated artefacts to be considered as creative output, the program that generated them must likewise be judged as somehow creative in its procedures. In Wiggins' (2006a) subsequent formalisation of Boden's model, the creative agent itself is bestowed with an evaluative function which it uses to assess its own output, effectively building a sense of creative value into the agent's procedure. Ritchie (2001) has likewise formally described the operation of creative systems in terms of an "inspiring set" of known good artefacts of a certain type: this set becomes both the basis for the way the system will structure its own output and the index beyond which the system must extend itself in order to be considered creative, in a process which involves sequences of self-evaluation moving from a basic set of possible items, through a consideration of the inspiring set, to the output of artefacts which are hopefully both new and valuable.

In more recent work, Ritchie (2007) considers the merits of the view that the creativity of a system should only be considered in terms of its output. Part of Ritchie's reasoning is that human creators are generally only judged on the basis of what they do, not how they do it. On the surface this might seem to be in line with Wiggins' definition of computational creativity in terms of "behaviour exhibited by natural and artificial systems, which would be deemed creative if exhibited by humans," (Wiggins, 2006b, p 210). In fact, though, it would be a mistake to take "behaviour" here in the Skinnerian sense of observable responses to stimuli; what is really in question in terms of behaviour is the way in which the agent goes about making the artefact. And finally, Gervás (2010) has proposed a model for creative output that involves cycles of production and reflection on the work in progress. This is again ostensibly in a similar vein to Ritchie's chain of evaluation of different stages in an overall creative process, but Gervás, in support of the significance of procedure, actually specifically suggests that it is perhaps misguided to try to build systems to appear operationally like creative humans, reasoning that there are a multitude of engineering solutions for a given objective, and blind imitation is rarely the best approach.

From these stances a range of approaches to evaluation emerge, aligned along two main axes: on the one hand, there is the problem of whether or not the system should be considered in terms of its internal workings, and on the other hand, there is the question of whether or not the system should attempt to be humanlike. But establishing what exactly counts as a creative process in the first place has proven extremely difficult. Where human creators are easily forgiven for keeping their methodologies secret – where, indeed, the mysteriousness of creativity is enshrined in humans through terms such as "genius" – such vagary is deemed unacceptable in a computer. The problem at least partially lies in the question of precisely where the act of meaning occurs: can a computational system really make meaning, or is it the observer who gives meaning to output which is merely the result of informational shuffling? In particular, a problem arises in terms of defining what counts as internal with regards to an information processing system. Given that the operation of a symbol manipulating machine is based on an interpretation of symbols which is fundamentally relative to a subjective observer (Putnam, 1988), the idea of a computational system being anything other than observable seems to fall apart, in which case everything that the actual system does can only be construed as output. If this is the case there is at least an argument to be made from a philosophical perspective for Ritchie's (2007) view that changes in the system's process must themselves be viewed as output in order to be assessed.

One practical approach to resolving these issues of evaluation has been formalised by Colton, Charnley, and Pease (2011), who, through their FACE model, propose a four step process for generating creative artefacts, or, in their terminology, expressions of concepts. Crucially, this process involves the establishment of framing information that potentially contextualises or justifies corresponding generative acts. The FACE model is complemented by the IDEA model, a framework specifically designed for the evaluation of creativity, both in terms of artefacts and actions. In an

explication of the theories behind these models, Pease and Colton (2011a) are motivated by an appreciation of the tensions that arise between creators and observers in the course of creative generation and evaluation, and seek to place the generation of new meaning in this dynamic relationship. By grounding the context of meaningful expression in public information, the hope is that the problem of trying to conceive of mechanical systems with internal states might be resolved.

The FACE model has been implemented by Colton, Goodwin, and Veale (2012): the output of the system developed by these authors offers, in conjunction with new poetry, a narrative alleging motivations on the part of the system in the course of poetic production. Furthermore, this narrative is grounded in an analysis of sentiments and concepts found in an external source, namely, in newspaper articles from a chosen date. This reflexive procedural commentary is specifically motivated by the view that observers do take into account creative process when evaluating an artefact, a stance which is also expressed by Colton (2008) in earlier theoretical work. By ascribing a phenomenology, however implausible, of intentions and emotions to the computational agent, the system generates a secondary level of artifice wherein the artefact is the result of some process of conceptualisation, representation, and execution. The hope is that humans will associate a capacity for creativity with the impression of intentionality.

What seems to be happening here is the simulation of precisely those properties of internal mental states that, as discussed in the previous section, have been attacked by contemporary cognitive scientists and philosophers of mind. Despite this, the stance taken in this paper is that this type of simulation is, broadly, the correct approach to take towards the evaluation of creativity—an evaluative act which, looking at it from the other end of the equation, might as easily be described as persuasion on the part of the agent. However, the stance here is also that mere mimicry of phenomenology is not ultimately a compelling argument for the creativity of a system. Rather, what is needed is a system that legitimately instantiates mechanisms with some similar properties to those that result in the appearance of mental states in cognitive agents.

In their zeal for non-representationalist, anti-dualist theories of mind, the contemporary mode of environmentally oriented approaches to cognition have arguably overshot the philosophical mark: not only do they reject the Cartesian stance; their rejection is so thorough that they neglect to properly consider why the mind/body divide has preoccupied Western thinkers for so long in the first place. But the appeal of the idea of an inner life of the mind is powerful on a collective level, running so deep in society that it has been instantiated in the form of intellectual property law, whereby authorship is ascribed on the basis of an ill defined "creative spark," (Fei, 1991). Indeed, in a legal sense, and therefore also to some degree on the scale of society, ownership of expressions is construed in terms of the distinction "between creation and discovery," (ibid). Elsewhere, McGregor (2014) has proposed that intellectual property law itself might be considered as one viable mechanism for the evaluation of creativity, and that something in the creative process or artefact might be offered up to appease the law's requirement for a distinguishing creative aspect. This is a problematic stance for the prevalent model of computational creativity, which, again per Boden (1990), involves a combinatorial exploration of a well defined state space, where the artefacts of such an exploration must be construed as discovered rather than generated. If the computational agent is to be presented as creative on a social level, then, it would seem the only course of action is to somehow trick the public into thinking of certain informational manipulations as being somehow inherently mental.

The idea of trickery isn't totally new to the field. In particular, where the theoretical work of Colton (2008), like that of Gervás (2010), plainly states that the computational agent should be straightforward about its own nature, the practical implementation of Colton, Goodwin, and Veale (2012) develops an agent that sets about selling its own product with an appeal to intentionality which might almost be described as deceptive. Similarly, albeit in a different domain, Leymarie and Tresset (2012) have designed an ingeniously conceived robotic portrait artist that is programmed to simulate behaviours the programmers have determined sitters and onlookers expect to witness in human artists: the robot enacts a roving quality to its video-camera eye, accompanied by built in pauses which create the illusion that the device is contemplating its work. The deception here, though, is transparent, and is committed with the good faith of honest artistry: it is unlikely that many observers believe these processes, which in the cases in question involves prefigured semantic networks and sentiment analysis or else an encoded parroting of creative behaviour, actually build up any kind of intentionality prior to the production of the output. Even a philosophically disengaged observer should not be expected to accept that phenomenology and intentionality can arise simply through the application of preconceived frequentist methods of data interpretation, or through the robotic rehearsal of a choreographed sequence of stereotypical gestures.

So it is proposed here that observers look for familiar processes when analysing creativity, but that this familiarity should be on the level of the impression of developed internal mental states rather than just superficial expressions; it is further proposed that the right approach to building creative agents is therefore to construct systems which remit the appearance of some kind of internal representations which are developed and manipulated in the course of searching for new, interesting artefacts in any given domain. The claim will be that, in such systems, while the base level of artefacts output for a target domain may be considered simply discovered within the search space chosen by the agent, creativity happens in terms of shaping the search space in the first place, not in terms of the subsequent traversal of that space, an idea which lines up nicely with Boden's (1990) notion of high level transformational creativity. Of course, this attempt to move creativity up a level, so to speak, suggests a secondary search space for new search spaces. Ritchie (2007) touched on this idea when he suggested that the creative process itself should be considered an abstract artefact

of the system, but what emerges is an infinite regression of spaces of spaces which immediately calls to mind the parallel homunculus problem in the philosophy of mind. This well travelled argument against representationalist theories of mind questions the basis for a secondary observation of mental representations by some internal observer – a homunculus, so to speak – an evidently necessary and likewise confounding condition for mind/matter dualism (see Dennett, 1991).

And this is precisely the point: entertaining this approach to the evaluation of computational creativity, namely, the consideration of an agent as being composed of a recursive hierarchy of creative search spaces, results in the same kind of untenable scenarios which characterise the dualist world view. The Cartesian outlook begs the question of who or what is perceiving internal mental states, and, more pointedly, suggests that these internal observations must likewise yield to some form of dualism, setting off a concatenation of ever deepening layers of internal states with no explanation of how this chain could terminate. In the same way, suggesting that a system becomes truly creative when it actually changes the parameters for discovering new and valuable artefacts necessitates a secondary search space with some sort of overview of the primary space from which it might seek the appropriate transformations; this secondary space, however, immediately becomes subject to the same criteria for transformation as the primary space, and an infinite regression rapidly develops. By this untenable mechanism of an infinite hierarchy of spaces in a finite system, a deeper operational correspondence between dualist theories of mind and transformational theories of computational creativity emerges. When the external impression of phenomenology is constructed by merely using information processing systems to analyse input and then combine indexical terms into the semblances of intentions and emotions, the impression of creativity can only be ephemeral. When a system actually reveals that it is operating in a way which establishes the kind of conceptual structures and recursive levels of abstraction associated with what is popularly, if erroneously, considered to be the dualist nature of cognition, on the other hand, the system has much more of a chance of being considered autonomously creative. The question, then, is what exactly qualifies as a semblance of dualist operation in a symbol manipulating system.

Implementing Representations

The next section of this paper will briefly consider two emerging computational models in terms of their potential as operational frameworks for computationally creative systems. Both vector space models and deep belief networks have been developed for the purpose of computing with high-level conceptual structures, and each system has been at least somewhat successful in its applications to specific informational domains. The question addressed here is whether the operation of either of these systems is such that they might be considered to produce the same kind of structures which observers imagine correspond to the mental representations attributed to the immaterial minds of humans under a dualist world view.

The hope is that these systems might show some promise in generating computationally graspable conceptual structures which can play a part in the act of meaning: more than just arrangements of data, these conceptual entities would stand for processes to be performed on data, abstract actions in the symbolic world of the computer, realised only through observation. The development of these systems for creative, generative purposes, however, is left for the future.

Vector Space Models

Initially developed as a mechanism for document indexing (Salton, Wong, and Yang, 1975), vector space models are built of high dimensional spaces whose dimensions correspond to the relational terms associated with a linguistic object: the object is described on the basis of the frequency with which each of the dimensional terms occurs in its context, and thus can be represented by a vector in the space. The idea is that similarity between two objects represented in such a space can be interpreted from the degree of the cosine angle between their corresponding vectors. In more recent work, vector space models have been applied to more basic problems of meaningfulness through distributional models of language, where words are represented in terms of their context and in particular through vectors representing either the frequency or the probability with which they occurred in the context of other words. This approach has been used to attack problems such as word disambiguation (Schütze, 1998) and compositional semantics (Mitchell and Lapata, 2008; Coecke, Sadrzadeh, and Clark, 2011).

The compositional approach in particular has revealed the utility of the mathematical nature of the vector space models. As illustrated in Grefenstette and Sadrzadeh's (2011) implementation of Coecke, Sadrzadeh, and Clark's (2011) framework, the properties of these kind of high dimensional representations allow for the composition of new representations through the use of Kronecker products, a technique which, by virtue of its non-commutativity, produces different spaces even for different combinations of the same words—a desirable outcome, given that word order can make a significant contribution to meaning in a sentence. This feature of vector space models allows for the construction of increasingly complex spaces as words are incrementally built into phrases and then sentences. The result is a system containing a vocabulary, so to speak, of highly modular compositional elements: the spaces of words can be easily concatenated into larger meaningful elements on the level of sentences, which become spaces themselves through the mathematical operations which can be performed on these types of structures.

In terms of computational creativity, what emerges from the perhaps somewhat complicated mathematics of vector space models is a mechanism for possibly representing what Davidson has described as "meanings as entities," (Davidson, 2001, p 116): the raw data of language become objects that can interact in ways that might produce valuable, surprising new semantic combinations. This approach to the composition of conceptual structures abstracts the problem of semantics away from the level of data processing, and likewise away from ungainly interventions of word associa-

tions and semantic ontologies that leave an observer wondering if the real creativity hasn't been imposed on the system through a preconceived framework. Instead, by generating and manipulating representations with operations that seem far removed from the logic of mental states or the syntax of the source language, a vector space is effectively promoted to the same level as the meaning-rich kind of encounter that humans have with the world and seems to thereby manifest some of the same mysteriousness associated with that way of being. Rather than relying on an externally grounded observation to give a system of symbols meaning, the objects that populate vector spaces can interact in ways native to their abstract mathematical domain, and in so doing instantiate entities that at least can be construed as conceptual representations analogous to the internal imagery of the Cartesian mental space.

While the use of vector space models for creative purposes remains unexplored, the indication from the work done in text analysis gives grounds for proposing that this could be a good method for likewise compositionally building linguistic artefacts which meet the constraints of a creative search space. And, importantly in terms of the subject of this paper, there seems to be good reason to hope that these conceptual structures might stand a chance of convincing a sceptical observer that a system employing them creatively could be utilising something similar to the types of internal representations which have been associated with the human use of language and the human mode of thought, per the likes of Chomsky and Halle (1968) and Fodor and Pylyshyn (1988).

Certain other systems have, in fact, taken a generative approach to vector spaces. The latent Dirichlet allocation model (Blei, Ng, and Jordan, 2003) is in particular a topic modelling technique that discovers topics within a range of documents and then builds a probability distribution for words across these topics. Latent Dirichlet allocation is generative in the sense that it picks potential words based on a probability distribution over a topic: the distribution of topics across a potential documents suggests likelihoods for the words which might occur in that document, albeit without the word ordering critical to a meaningful use of language. This is not necessarily an ideal strategy for modelling creative behaviour, however, as, in addition to the absence of compositionality, generative models tend to predict output that is highly likely but, conversely, not very surprising. In the context of generating meaningful and unexpected new language, the compositional approach discussed above seems to hold more promise for finding the semantically loaded output expected from a creative agent.

Deep Belief Networks

Where vector space models have proved particularly powerful for language, deep belief networks have been used effectively for work in the domains of both computational linguistics and computer vision. Deep belief networks were proposed by Hinton, Osindero, and Teh (2006) as high parameter frameworks that would learn to identify handwritten numerals by developing a model for generating the same artefacts. In this case the generative quality of deep belief networks do specifically point to a creative application, in that the network actually learns to match new, noisy percepts with semantically tagged representations by actually learning to produce those representations in an initial stage of development. Across its many levels of processing, the network purportedly develops different layers of feature detection, and these features – for instance, lines, contours, or, eventually, at a high level, concepts – arguably convey the impression of the internal states corresponding to the mental perception of properties in the world.

The idea is that densely connected networks consisting of a large number of artificial neurons rising over several diminishing layers in a pyramid type structure can be efficiently and effectively trained if they are constructed with the right kind of architecture. The keys to this architecture are a special mechanism at the low level related to Ackley, Hinton, and Sejnowski's (1985) earlier work on Boltzmann Machines (another type of neural net that utilises a stochastic mechanism), as well as the simplicity with which the connecting weights between neurons are updated. With their highly interconnected structure, deep belief networks might be seen as the next phase in the historic cycle of interest in connectionist approaches to computing; the new element in this latest manifestation is the stacking of several operational layers where parameters are established in a layer by layer fashion.

The operational key to deep belief networks is the idea that, by allowing a single neuron on a higher level to represent the clusters of neurons which feed into it from the level below, an exponential reduction of computational space can be realised (Bengio, 2009). In this way, these networks establish elevating levels of abstraction that might be construed as internal representations. Indeed, in precisely this sense, deep belief networks seem to relate to the idea of the act of meaning by which potentially diffuse visual data are resolved into higher level percepts with some semantic value. The argument put forward here is that this on the one hand instantiates the approach to cognition through the creative reconstruction of anticipated events in the world endorsed by Wiggins (2012), and, on the other hand, creates structures which might be recognised by an observer as something similar to internal mental states.

Going back to some of the original literature from the first wave of neural networks, the structure of the human brain was clearly a primary motivation in the effort to compute using weighted networks of nodes (McCulloch and Pitts, 1990). Deep belief networks have inherited this property, and have taken inspiration from another aspect of neuroscience: the multiple layers in a deep belief network specifically resemble the hierarchical structure of the visual cortex in the human brain (Bengio, 2009). Indeed, Serre et al. (2005) have done work towards isolating the ways in which different levels of the primate visual cortex build up different aspects of representations of raw visual stimuli, ultimately resulting in the high level perceptions of parametrically bound entities which seeing, thinking agents experience in the world. In the same way, deep belief networks seek to use increasingly complex clusterings of input data to form higher levels of representation within their architec-

ture. Coupled with the fact that these systems are fundamentally generational, such networks seem like an excellent candidate for consideration as visually creative agents with a convincing impression of internal representations, and probably warrant exploration in other domains, as well.

Conclusion

Dualism was born of a simple thought experiment: Descartes (1911) imagined himself plagued by a demon fixated on deceiving him, and in response strove to strip away from his experience of reality everything which could possibly be considered illusory. He was left with the certainty of his own irreducible mental existence, but maintained that this existence must also be involved in some sort of likewise irreducible physical reality. Since even before Descartes' time, various similar imaginative exercises have characterised the development of Western philosophy, from Plato's (1892) cave to Wittgenstein's (1967) beetle. Notable recent thought experiments seeking insight into the mind have included Putnam's (1996) twin earth, Davidson's (1987) swampman, and Chalmers' (1996) philosophical zombies—and, as mentioned earlier, the computer has played a part in some other recent enquiries, though generally as a device for demonstrating the absurdity of certain views of cognition that can be reduced to mere data shifting.

The purpose of pointing out this tradition of thought experiments is to highlight the role which the peculiar act of introspection has played in the development of modern Western philosophy. The preoccupation with intentionality and phenomenology have grown out of an intellectual culture of examining the self, and the willingness which humans have to accept the creativity and indeed the very meaningfulness of the expressions of other humans seems to stem from the recognition of a similarly calibrated other-self. What has been proposed in this paper is that the external alienation of an encounter with a computational system can be replaced with a look into the exposed operations of the system, and, in this exposure, there may be some hope of acceptance that the symbol manipulating machine is behaving in a way which is creative, in the operational sense of behaviour described by Wiggins (2006b).

The idea that information processing systems should be investigated for indications of familiar processes in order to be considered creative is not new. Gervás (2010) has argued that hardware which operates in a highly parallel manner should be taken more seriously as a candidate for instantiating creative agency, as this type of procedure to some degree mirrors the evident dispersion of activity in the human brain. Perhaps even more fundamentally, Pease, Winterstein, and Colton (2001) call for a criterion of procedural complexity intended to measure the extent of the creative search space and the difficulty of the agent's traversal of this space. It is not clear, however, how such mechanisms don't become just another aspect of the agent's output, adjunct to the creative artefact itself. What is called for in this paper is a probing of the machine – an extrospection, so to speak – for the representational type of processes that society at large seems to deem, in the tradition of Descartes, should count as cognitive and potentially creative. It is for the observer to seek out and identify the structures which form these representations rather than for the system to simply present them either though a statement of intentionality or an exposure of process.

What form these representional structures would take remains to be defined, though two possible candidates have been proposed here. A further area for enquiry is the question of what kind of observer would be able to recognise these structures in the first place: is some combination of expertise in philosophy and computer science necessary in order for a computationally creative agent to be recognised as such? Ideas along these lines have been proposed by Pease and Colton (2011b) and Boden (2014), all of whom suggest that computational creativity may be best judged by an audience with a degree of knowledge about how computers work. On the one hand, the idea of expert criticism informing the public as to the value of creativity has long been common in various domains such as art, literature, and film, and some degree of expertise is probably necessary to achieve recognition of the relatively complex frameworks discussed earlier in this paper. On the other hand, relying on computer scientists for assurances of the legitimacy of creative agents risks further alienating an audience already confronted with a very new and different mode of creation, and, indeed, of creator.

So even the proposal for a solution to the problems laid out in this paper seems to open the door on another potential debate. Such is the nature of philosophy. Nonetheless, this paper has sought to show that computational creativity as a field is an appropriate platform for engaging in discussions about not only aesthetics but also cognition and theories of mind, and has at least presented an avenue for further philosophical investigation.

Acknowledgements

This research has been supported by EPSRC grant EP/L50483X/1.

References

Ackley, D. H.; Hinton, G. E.; and Sejnowski, T. J. 1985. A learning algorithm for Boltzmann machines. *Cognitive Science* 9(1):147–169.

Bengio, Y. 2009. Learning deep architecture for AI. *Machine Learning* 2(1):1–127.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Boden, M. A. 1990. *The Creative Mind: Myths and Mechanisms*. London: Weidenfeld and Nicolson.

Boden, M. A. 2006. *Mind as Machine: A History of Cognitive Science*. Oxford: Clarendon.

Boden, M. A. 2014. Skills and the appreciation of computer art. In *Proceedings of AISB14CC*.

Chalmers, D. J. 1996. *The Conscious Mind*. Oxford University Press.

Chemero, A. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: The MIT Press.

Chomsky, N., and Halle, M. 1968. *The Sound Pattern of English*. New York: Harper and Row.

Coecke, B.; Sadrzadeh, M.; and Clark, S. 2011. Mathematical foundations for a compositional distributed model of meaning. *Linguistic Analysis* 36(1-4):345–384.

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational creativity theory: The FACE and IDEA models. In *Proceedings of the International Conference on Computational Creativity*.

Colton, S.; Goodwin, J.; and Veale, T. 2012. Full-FACE poetry generation. *Proceedings of the Third International Conference on Computational Creativity* 95–102.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of the AAAI Spring Symposium on Creative Intelligent Systems*.

Davidson, D. 1987. Knowing one's own mind. In *Proceedings and Addresses of the American Philosophical Association*, volume 60, 441–458.

Davidson, D. 2001. Truth and meaning. In Martinich, A. P., ed., *The Philosophy of Language*. 114–124.

Dennett, D. C., and Searle, J. R. 1995. 'The Mystery of Consciousness': An Exchange. *The New York Review of Books* 42(20).

Dennett, D. C. 1991. *Consciousness Explained*. London: The Penguin Press.

Descartes, R. 1911. *The Philosophical Works of Descartes*. Cambridge University Press. Translated by Elizabeth S. Haldane.

1991. Feist Publications Inc. v. Rural Tel. Service Co., 499 U.S. 330.

Fodor, J. A., and Pylyshyn, Z. W. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1-2):3–71.

Gervás, P. 2010. Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, 23–30.

Gibson, J. J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Miffline.

Grefenstette, E., and Sadrzadeh, M. 2011. Experimental support for a categorical compositional distributional model of meaning. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Hinton, G. E.; Osindero, S.; and Teh, Y.-W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18(7):1527–1554.

Leymarie, F. F., and Tresset, P. 2012. Robot drawing and human engagement. In *Proceedings of the 5th International Workshop on Human-Friendly Robotics*.

McCulloch, W. S., and Pitts, W. H. 1990. A logical calculus of the ideas immanent in nervous activity. In Boden, M. A., ed., *The Philosophy of Artificial Intelligence*. Oxford University Press.

McGregor, S. 2014. Considering the law as an evaluative mechanism for computational creativity. In *Proceedings of AISB14CC*.

Mitchell, J., and Lapata, M. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08:HLT*, 236–244.

Newell, A., and Simon, H. A. 1990. Computer science as empirical enquiry: Symbols and search. In Boden, M. A., ed., *The Philosophy of Artificial Intelligence*. Oxford University Press. 105–132.

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

Pattee, H. H. 2008. Physical and functional conditions for symbols, codes, and languages. *Biosemiotics* 1(2):147–168.

Pease, A., and Colton, S. 2011a. Computaional creativity theory: Inspirations behind the FACE and IDEA models. In *Proceedings of the International Conference on Computational Creativity*.

Pease, A., and Colton, S. 2011b. On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In *Proceedings of the AISB symposium on AI and Philosophy*.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating machine creativity. In *Proceedings of ICCBR-2001*.

Plato. 1892. *The Republic*. Oxford University Press.

Putnam, H. 1988. *Representations and Reality*. MIT Press.

Putnam, H. 1996. The meaning of "meaning". In Pessin, A., and Goldberg, S., eds., *The Twin Earth Chronicles: Twenty Years of Reflections on Hilary Putnam's "The Meaning of 'Meaning"'*. Armonk, NY: M.E. Sharpe. 3–52.

Ritchie, G. 2001. Assessing creativity. In *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Rowlands, M. 2010. *The New Science of the Mind*. Cambridge, MA: The MIT Press.

Salton, G.; Wong, A.; and Yang, C. S. 1975. A vector space model for automatic indexing. In *Proceedings of the 12th ACM SIGIR Conference*, 137–150.

Schütze, H. 1998. Automatic word sense discrimination. *Computational Linguistics* 24(1):97–123.

Searle, J. R. 1990. Minds, brains, and programs. In Boden, M. A., ed., *The Philosophy of Artificial Intelligence*. Oxford University Press. 67–88.

Serre, T.; Kouh, M.; Cadieu, C.; Knoblich, U.; Kreiman, G.; and Poggio, T. 2005. A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, MIT Computer Science and Artificial Intelligence Laboratory.

Sloman, A. 1978. *The Computer Revolution in Philosophy: Philosophy Science and Models of Mind*. The Harvester Press.

Varela, F. J.; Thompson, E.; and Rosch, E. 1991. *The Embodied Mind*. Cambridge, MA: The MIT Press.

Wiggins, G. A. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19:449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Generation Computing* 24:209–222.

Wiggins, G. A. 2012. The mind's chorus: Creativity before consciousness. *Cognitive Computing* (4):306–319.

Wittgenstein, L. 1967. *Philosophical Investigations*. Oxford: Basil Blackwell, 3rd edition. trans. G. E. M. Anscombe.