Computational Creativity and Consciousness: Framing, Fiction and Fraud
Paper type: Study Paper

Geraint A. Wiggins
AI Lab, Vrije Universiteit Brussel, Belgium &
EECS, Queen Mary University of London, UK
geraint@ai.vub.ac.be

Abstract

Computational Creativity, like its parent, Artificial Intelligence, suffers from ill-definition: both "creativity" and "intelligence" are difficult, perhaps even impossible, to define. Both fields have also suffered from confusion about the relationship between their key concept and the equally problematic concept of "consciousness". Computational Creativity, however, has yet to address this issue effectively, which can only be detrimental to the field. This paper attempts to lay out the issues, to identify useful boundaries, particularly with respect to framing and the generation of meaning, and to map out ways in which Computational Creativity research may navigate the landscape of scientific possibility, while remaining true to itself.

Overview

In the current paper, I attempt to engage with some open questions regarding the relation of creativity, and specifically computational creativity, to consciousness and meaning. While definitive answers are not yet achievable, I suggest that the discussion is progressive and useful. In this context, I discuss computational creativity research methods, with particular emphasis on the notion of *framing* (Pease and Colton, 2011). First, relevant aspects of the philosophy of artificial intelligence (AI) are summarised, pointing out a new interpretation. Then, aspects of research on consciousness and its relation to AI are reviewed. While most of the ideas presented in these first two sections are probably familiar to many readers, they supply a specific context for what follows. Next, the relevance to computational creativity, practical and theoretical, of these issues, is discussed, in context of recent publications in the field. Finally, some principles are proposed that may help computational creativity research to make progress as a scientific endeavour.

In summary, the conclusions argued in this paper are as follows:

1. Computational creativity, as a scientific discipline, should (and mostly does) primarily focus on identifying those elements of creative systems that are *necessary* to allow creativity to emerge from their operation. Thus, computational creativity can predominantly be viewed as an *ex post* phenomenon *emergent from* computational systems.
2. Any attempt to introduce *aesthetics* into computationally creative systems must account for the origin of the aesthetic so introduced in a philosophically sound way.
3. In order to demonstrate scientifically that computational systems enjoy such an *ex post* capacity as creativity, evaluation methods used must be *honest*, in the sense that any detected emergence of creativity should be, whether directly or indirectly, explicitly attributable to the processes within those systems. In particular, when using metaphor to explain outcomes from computational creativity (or any other research field), it is important to be explicit about the metaphorical nature of the explanation.
4. This notion of honesty must extend into the application of framing to computationally creative systems and their outputs.

Context: AI, Creativity, and Consciousness

In his seminal paper, *Computing Machinery and Intelligence*, Alan Turing (1950) sets out to address the question "Can machines think?" He begins from the position that, stated in this way, the question is "too meaningless to deserve discussion" (p. 442), because the words "machine" and "think" are ill-defined. To address this problem, he defines precisely what he means by "machine"—a digital computer—and proposes a thought experiment to help understand the question, the Imitation Game, which has come to be known as "The Turing Test". We are invited to consider the question of whether a machine can convince a human that it is a human, in competition with a human, who is attempting the same, via a typed dialogue.[1]

The status of "intelligence" in Turing's extraordinary paper is never made specific. Aside from the title, the word occurs only once, explicitly in the context of *human* intelligence, and the connection between "can…think" and "is intelligent" is left for the reader to assume. The word "intellect", also, is used only to describe *human* capacity. So Turing is proposing a proxy for the question, "can machines think?", which avoids the question of what it means to "think", by substituting a comparison with something (a healthy adult human) that is generally agreed to be able to

[1] When people appeal to the "The Turing Test" in the modern scientific literature, it is usually to a less challenging form, in which there is no opponent, and often even no dialogue.

think. Many definitions of Artificial Intelligence are similarly couched in terms of "things a human can or might do" (e.g., Bellman, 1978; Haugeland, 1985; Charniak and McDermott, 1985; Winston, 1992), and in some definitions of computational creativity (e.g., Wiggins, 2006a).

Turing addresses various objections to his replacement of that original question that are not relevant to the current discussion: religious dualism, pro-human and anti-machine arguments of various degrees and kinds, and formal mathematical arguments. Two further counter-arguments, that are relevant here, are what Turing calls "the argument from consciousness" (p. 445) and (charmingly) "Lady Lovelace's Objection" (p. 450). These are addressed in later sections.

It is noteworthy that Turing never explicitly suggests that the Imitation Game should be used as a scientific test for intelligence in a given computer—instead, he explicitly posits it as an alternative to the general *question* about "imaginable" machines (p. 436), and not about a particular machine. He refers to the (computational) Imitation Game as a "test" only three times. Nevertheless, the paper is widely supposed to be a specific proposal for a functional test to identify the presence of intelligence. An alternative view of Turing's argument, since he also does not claim that a computer is like a human brain[2], is as follows. Turing *could* be presenting a thought experiment, making the point that, without a *functional* definition of "thinking", one cannot distinguish "real thinking", done by a person, from the outward appearance of "thinking", by an adequately programmed machine, even if what the machine is doing is not "real thinking" at all. The inevitable conclusion from this perspective is: if the behaviour of the machine *appears* close enough to behaviour arising from human "thinking", one can no longer tell whether it is "really thinking" or not. This argument bears comparison with the "process *vs.* product" argument in creativity studies, and the answer is the same: for a meaningful comparison between human and artificial, *both* process (i.e., mechanism) *and* product are important.

This interpretation is not the widely accepted intent of Turing's paper. However, whether or not it is his intended interpretation is immaterial for the purposes of the current argument. For, whether Turing intended it or not, the above is indeed a valid consequence of his argument: if we define intelligence only by our ability to recognise its effects, we can be fooled. In principle, a sufficiently detailed, but nevertheless stateless (as defined by Russell and Norvig, 1995), agent, **A**, will be enough to fool us.[3] This is a version of the Chinese Room argument (Searle, 1980), to which we return below.

In summary, the best definition available of intelligence as a property of an agent renders that property detectable only by observation of behaviour of that agent in a given context: a firmly *ex post* definition. Therefore, two healthy adult humans, sitting, motionless, side-by-side are utterly inscrutable and indistinguishable with respect to that property, no matter who they are.

Russell and Norvig (1995), who currently inform students worldwide, effectively circumvent this problem by defining it away: for them, artificial intelligence should exhibit rational behaviour, which is in turn defined in terms of actions that lead to a goal. Thus, artificial intelligence is rational agency, and an agent that does not have goals (like the stateless agent, **A**, above) is by definition not intelligent, reversing the logical order to *ex ante*. While this definition works well from the practical engineering perspective of getting at least something done, it quickly becomes clear that it is incomplete, because intelligent organisms do more than just *attain* goals—they also *identify and formulate* them. Thus, Turing's "thinking" machines are reduced to reactive (albeit flexible) slaves whose goals are dictated, presumably by proactive humans. Furthermore, this particular definition of "rationality" excludes what, in a human, would be considered emotion or affect—that is, the *aesthetic* side of intelligence. In defence of Russell and Norvig (1995): one has to start somewhere.

Lady Lovelace's Objection Ada, Countess of Lovelace was arguably the first software analyst, studying the output from Babbage's computing engines in the first half of the 19th Century. Startlingly ahead of her time, she considered the possibility of computational creativity, but concluded that a machine was not capable of creating something new, because it could only do what its program(mer) told it to do (Countess of Lovelace, 1842). Computational creativity researchers remain under siege by this counter-argument, 200 years later: the Objection has been resurrected at every single public talk on computational creativity that the current author has given in the past 20 years. Turing (1950, p. 450), equally startlingly ahead of his time, refutes it with an argument that still holds: if a machine can learn, and base its productions on what it has learned, then there is no reason why, in principle, those productions cannot be novel, and independent from the programmer. This is a large step towards autonomous creativity.

The Argument from Consciousness Turing responds to "The Argument from Consciousness", made by Jefferson (1949), that a machine, capable of "real thinking", should be able to perform creative acts because of the thoughts and emotions that it *feels*, and also know that it performed them. Thus, human-level self-awareness is invoked, in what is often classed as the highest level of consciousness (Merker, 2007, and below): experiencing not only one's existence, but also awareness of one's own existence; experiencing knowing that one exists; knowing that one is capable of action in the world, and so on. In his rebuttal, Turing does not draw a clear line between this construct and what he means by intelligence. He argues that if consciousness is a necessary part of thinking, then the only way to demonstrate true thinking by a machine would be to be the machine and to experience the thinking, first hand: the solipsist position. We avoid the solipsist position in human society, Turing charmingly notes, by agreeing "the polite convention that everyone

[2] Though he did, in fact, colloquially refer to Universal Turing Machines as "Brains" (Hodges, 1992).

[3] A reasonable rebuttal here is that such an agent would be impossible to build, because there would be too many cases to include in its production list. However, for the purposes of the current argument, the theoretical possibility suffices, because, no matter how complex the world, the Imitation Game endures only for finite time.

thinks" (Turing, 1950, p. 447). Thus, the answer seems to be "this question is ineffable" and that people will yield before being cornered into solipsism. This is a rhetorical argument, not a logical one, and so it does not satisfy.

More usefully, Turing confirms that he does see the "mystery about consciousness" (Turing, 1950, p. 447), and denies the need to solve that mystery before answering his central question. In doing so, he implies that consciousness is something (at least partly) separable from intelligence, and this is more telling than his actual rebuttal: if intelligence can be reproduced in a machine without consciousness, then intelligence sits far more comfortably in what Turing explicitly defined as a machine than otherwise.

The Chinese Room Perhaps surprisingly, since Turing defuses this issue explicitly in his paper (see above), the Chinese Room began as an argument against what Searle (1980) called "strong AI": that a machine could have a mind in exactly the same sense that (we all politely assume) humans do. This definition tacitly conflates the property of intelligence with the property of consciousness, and these two are not the same (Turing, 1950; Preston and Bishop, 2002). Briefly summarised, the argument posits a closed room with a person inside, who does not speak or read Chinese. That person has access to a large body of knowledge, expressed in terms of Chinese characters. This knowledge is presented in such a way that it may be deployed by having the person match the characters together, without understanding their meaning. Thus, questions, posed in Chinese, and posted on paper through a hatch in the wall, maybe answered, also in Chinese, by the person matching and copying the relevant characters, and passing them out on paper through another hatch. Thus, the room appears to answer questions, but it cannot be said to *understand* the questions and answers in the sense that the person within would understand them in the person's own language. Searle (1999) changed his position on this, acknowledging that the Chinese Room is not so much an argument against artificial intelligence as against machine consciousness, and in this context it is indeed more successful: few artificial intelligence researchers claim to be developing human-like consciousness in their machines. Indeed, such an attempt would be ethically questionable, because, if it were successful, the "off" switch would become a murder weapon.

Context: Conscious Experience and Aesthetics

Previously, we rehearsed the Chinese Room argument regarding consciousness in computers. We now address the concept of conscious experience in general, as discussed by many philosophers and others (e.g. Wittgenstein, 1958; Nagel, 1974, 1986; Searle, 1980, 1999; Dennett, 1991; Merker, 2007, 2013; Shanahan, 2010). Although we are ultimately interested in the relationship between computational creativity and consciousness, we begin with the thought experiment of Nagel (1974): "What is it like to be a bat?"

First, it is important to understand that Nagel's usage of "What is it like to be…" is more specific than the everyday English usage. The author of the current paper, a university professor, could imagine what it is like to be, for example, a politician, thinking through the visible aspects of a politician's life, and transferring human perspectives between life experiences, real and imagined. This is the everyday usage: the professor is imagining what the politician's life is like from her own perspective. Nagel's usage is more specific: he refers to the bat's experience of its own existence, explicitly ruling out the experience of a human *pretending* to be a bat[4]; this will be significant later in our argument. It is like something to be the current author; it is like something *else* to be another human, the reader perhaps, though experience suggests that there are commonalities which can be conventionally agreed via reference to the external world; it is like something *else again* to be a bat, and there are aspects of bat experience that are not communicable to humans, and therefore could not be conventionally agreed, even if we could discuss them.

For the purpose of the current discussion, it is helpful to carry Nagel's argument further. The reader is now invited to try to imagine, in Nagel's sense, *what it is like to be a rock*. Even though the behaviour of rocks is significantly less complicated than that of bats, it is impossible for a human to imagine what it is like to be a rock. This is not just because of the significant physical differences between rocks and humans, but because, in Nagel's sense, *it is not like anything to be a rock*. We cannot meaningfully say "from the rock's perspective," because the rock *has* no perspective. We cannot imagine what it is like to be a rock, from the rock's perspective, any more than we can imagine our own future experience of being dead.

Having established an experiential boundary between the rock ("not like anything") and higher biological species ("like something"), we skip discussion of fungi, plants, etc., to focus instead on computers. Some of the components of a computer are made of rocks. For the most part, components of a computer are like rocks, in that they they exist, to all intents and purposes, statically. When an electrical current is applied by an external agency, the chemical construction of some of those very small pieces of rock changes, and electromagnetic states are manipulated in such a way as to correspond with meanings imposed by an external viewer. Computers are designed to be efficient, in the sense that their operations are performed with the minimum of energy waste, doing only what has to be done, and nothing else. As Turing (1950) notes, their electrical currents form a commonality with the brain, but this is too weak a commonality to suggest that a computer works like a brain, or that it has the properties of a brain. For a clearer counter-example, consider the world-wide telephone network, whose entire business is the transmission of electrical signals: we do not conclude that it is functionally like a brain.

It is sometimes proposed that if a computer large and complex enough were built, then it would be conscious. However, when asked *why* this would be, proponents of the idea

[4] Nagel chose the bat for his example because it is a higher animal with very significantly different sensory capabilities from a human. Thus, people do not find it difficult to attribute consciousness to a bat, but they do generally find it difficult to imagine, for example, how the bat experiences its echo-location system.

have no mechanistic answer. Therefore this solution must be rejected along with other non-scientific proposals and fables of the supernatural.

What matters is that *there exists no evidence whatever that any entity exists, made of the same materials and working by the same processes as a computer, that is conscious*. On the contrary, the available evidence suggests that *only* living, biological entities are conscious, except in exotic definitions of the concept (e.g., Tononi, 2004), which are highly contested (e.g., Merker, Williford, and Rudrauf, 2021). Absence of evidence is not evidence of absence; however, since no known non-living example exists of Nagel's kind of consciousness, a scientist wishing to propose consciousness in non-living circumstances must provide an account of how, or at least where, it arises, in order to be philosophically convincing.

A categorical range of consciousness is expressed in the Indian "scale of sentience" (Merker, 2007):

"This"
"This is so"
"I am affected by this which is so"
"So this is I who am affected by this which is so"

Note that this is a scale, not of magnitude, but of ordered categories of successively superordinate kind. We can place humans in the fourth category. We speculate that dogs, for which evidence of self-awareness is lacking, but which seem capable of feeling how the environment affects them, might be in the third category. Lower molluscs (e.g., mussels) might be in the first, while some cephalopods (e.g., octopuses) are certainly in the third or even the fourth category. Computers and rocks, however, are not in any of these categories, *according to any extant evidence*.

What it is like to be a conscious entity includes what it is like to experience the environment of that entity: this is the first level of sentience, above. The instances of such experience, *qualia*, are themselves contentious (cf. Dennett, 1991), and they are unverifiable because they are *private*, in the strong philosophical sense: they are not directly communicable. To see this, consider the colour called "blue". One person may use the word "blue" to signify (part of) the experience of seeing a particular object to another person, and the second person may agree that the label "blue" is appropriate for that object. But it is unknowable whether the two people experience the same thing. For the avoidance of doubt, the question lies not in the external stimulus, because the light reaching the eyes of the two individuals may be measurably the same, but in the internal, private response to the stimulus of the two viewers. To introduce unusual terminology: there is a *feeling of blue*, private to each individual, which is experienced simultaneously, and labelled with a common word.

Computers and the Feeling of Meaning

Now compare our two humans with a computer equipped with an RGB raster camera. The camera can measure the light reflected from the same scene viewed by the humans, above, and upload a corresponding matrix to the computer. The computer's processor can compare these measured numbers with a range, and output the symbol "blue" when a programmer has deemed this appropriate, or when it has learned the necessary association from data. In spite of the fact that the computer seems to have grasped language, a capacity of only the fourth category of sentient beings, it has merely *measured* blue, lacking any mechanism with which to *feel* it[5]; it has not even reached the first level of sentience. The *anthropomorphic illusion* thus produced is beguiling, and will be important later in the current argument. That illusion permeates popular discourse: we talk metaphorically about "what the computer thinks" and "knows", when what we really mean is that some numbers have been used to represent some information by a programmer. Humans use metaphor to allow us to discuss things whose detail we do not know; doing so is an important part of communication, since it allows us to learn as we converse, by filling in the gaps as we go along. But it has a down-side: it can lead us to attribute capabilities or properties to the target of the metaphor that are incorrect and misleading.

The association by humans of the word "blue" with the relevant *quale* may also be viewed as defining the meaning of the word, "blue". One can give more objective measures, based on frequency of light, but this is misleading, since human word usage long predates such possibilities. Thus, the memory of the feeling of blue is the meaning of "blue" for each person who has seen a blue thing, and who uses the word. The word "qualia" is conventionally reserved for sensory experience, such as this. We posit that it is reasonable to extend the concept of sensory feelings, as exemplified here, into more abstract knowledge. Consider, for example, the idea of small, integer numbers. To think of the abstract concept of "two", a human need not see two things, nor hold up two fingers, nor count up from zero; further, the concept of "two" does not exist independently in the world, but only as a relation between things that exist in it. Nevertheless, humans have a feeling about what "two" means, which may be expressed in arithmetic, but which can still be felt without such expression, exactly as the reader did when considering the "blue" example, above. The contention here is that meaning is a construction of consciousness in context of memory, which begins with qualia building blocks, and which can assemble arbitrarily complex entities from smaller ones. Those entities acquire private meaning by virtue of their very existence, based on that of their components. To make them public, humans must realise them, either literally by building them, or via sequences of word labels that describe them, or via other more precise descriptions, such as mathematical specifications. Such *feeling/meaning*, I propose, is the fundamental stuff of human thinking and therefore of human creative thinking.

Somewhere on the constructive continuum between qualia and philosophical concepts such as "qualia", lie the feeling/meanings that drive biological organisms, some of which are close to, but not the same as, direct sensory experience, and many of which arise from sensing internal states of the body: for example, hunger is the feeling/meaning induced

[5] Of course, a dualist approach to human consciousness may be disenterred here, and applied to computers. We eschew such superstitious or mystical notions.

by metabolism of glycogen in the liver. More complex are what are generally called *aesthetic responses*: the feeling/meanings that result from often complex and extended experiences such as music, literature or visual art, and which also arise in humans involved in science, engineering and mathematics, in perhaps less obvious ways. In summary, experience produces feeling/meaning, and knowledge labels feeling/meaning or constructs new feeling/meaning from existing components.

It is for this reason that the research field of computational aesthetics (Hoenig, 2005) concerns itself with the *simulation of human aesthetics* and not with the development of aesthetics that might be felt by a computer; similar approaches appear in computational creativity (e.g., Norton, Heath, and Ventura, 2013). It is not like anything to be a computer; a computer does not feel, as there is literally no place for qualia. Therefore, a computer cannot have aesthetic experiences, which are, for humans, internally, nothing but feeling/meaning. Below, this will have consequences for the conduct of computational creativity research.

The Nature of Computational Creativity

Hodson (2017, p. 3) suggests that computational creativity, as a research field, has committed a "fundamental misunderstanding" by assuming that creativity is an *ex ante* phenomenon, rather than an *ex post* phenomenon. In other words, he writes, the field presupposes that creativity is caused by certain cognitive processes, rather than being something that is observable in context when it happens, no matter how it is produced. This interpretation is generally incorrect. Rather, in computational creativity, many, or even most, researchers seek systems that are imbued with the properties that *may afford* novel and valued outputs, which may only subsequently—i.e., *ex post*—be judged creative. In other words, we seek systems with the capacity to be "deemed creative by an unbiased observer" (Colton and Wiggins, 2012). Increasingly much research is devoted to the evaluation of systems *qua* creative systems (e.g., Ritchie, 2007; Jordanous, 2012; Wiggins et al., 2015; Jordanous, 2018). In other words still, we seek to understand what is *necessary* in order to achieve creativity (identified *ex post*), while not expecting to find what would be *sufficient* (*ex ante*) for creativity within a given creator. A very clear example of this approach may be found in the Empirical Criteria for Attributing Creativity to a Computer Program (Ritchie, 2007).

The *ex post* nature of creativity brings with it a danger—the same danger associated with Turing's Imitation Game, above: is a creative system *actually* creative, or is it merely *appearing to be* creative? In an artistic context, this distinction may be viewed as unimportant, or may even be itself the subject of artistic question. In a scientific context, it is problematic, particularly in the context of framing (Pease and Colton, 2011), to which we return below. Here, the gap between *what is sufficient for the perception of creativity* and *what is necessary for creativity* will be important.

Returning to the topic of this section: recall the differences between computational creativity and traditional AI. In particular, computational creativity broadens the scope of the intelligent behaviour that it studies beyond that which can be modelled as mere exhaustive search through combinations of solutions to a problem afforded by a representation that is specifically designed for that purpose. Here, the conceptual space of Boden (2004) is paramount, and so is her notion of *transformational creativity*. This is an operation which changes the very search space, a capacity far beyond traditional artificial intelligence systems. The requirement for a value function which is *not* restricted to a pre-defined meta-level measure of the pre-established solution space (Algorithm A* uses such a restricted heuristic)[6], and for the capacity to exit the conceptual space while still generating valued ideas, modelled as *aberration* in the Creative Systems Framework (Wiggins, 2006a,b), distinguish this view of computational creativity from traditional AI. Recalling that any algorithm may be written as a search algorithm, the difference might be crudely stated, thus: AI searches for solutions to problems, knowing that the solutions may theoretically exist but not where they are, while computational creativity searches for novel ideas of a particular form, that it values, but does not expect.

Aesthetics in Computational Creativity

Notwithstanding the current, and appropriate, emphasis on evaluation in computational creativity, there are still points where its philosophy becomes moot. It seems generally agreed that *internal evaluation* in a computational creative system corresponds with the aesthetic response of a human creator to her own work. Thus, that evaluation is a simulation of, or substitute for, the human creator's feeling/meaning. The same is true in some computational creativity systems. For example, the computational artist, DARCI (Norton, Heath, and Ventura, 2013), derives aesthetic labellings for images by learning from descriptions made by humans.

Colton (2019, §5) develops a roadmap for computational creativity, in seven steps. At level 2, the level of "appreciative systems", a creative system designer must "encode [their] aesthetic preferences into a fitness function"; this, we suggest, is slightly less than DARCI's learning capability. In level 3, the level of "artistic systems", a creative system designer must "give the software the ability to invent its own aesthetic fitness functions and use them to filter and rank the images that it generates." The contrast is clear: at level 3, human aesthetics are out, and machine aesthetics are in. (In parenthesis: the notion of generate-and-test appears to dominate here; whereas it is to be desired that advanced creative systems would not be restricted to that approach, but be rather more deliberate in the construction of their outputs. It seems unlikely, however, that Colton strongly proposes that all "artistic systems" should be limited to generate-and-test.)

The problem at level 3 is that there is no discussion of what it means for a machine to have an aesthetic. Given the absence of feeling/meaning in a machine, as argued above, the phrase "machine aesthetic" becomes a contradiction in terms, and therefore meaningless. Perhaps Colton requires his computers to be capable of feeling/meaning? But later he and colleagues say this is not the case (Colton et al., 2020).

[6] Computational creativity is not the only subfield breaking these chains: meta-heuristics research asks some of the same questions.

Perhaps he intends a sort of arbitrarily generated selector imposing an arbitrary choice, unrelated to feeling/meaningful aesthetic response? But this, therefore, should not be called "aesthetic".[7] There is no mention of co-creativity (e.g., Kantosalo and Toivonen, 2016) here, and anyway, such collaboration with a human would lead the aesthetic function back towards (if not directly to) something that models human aesthetics, which Colton has rejected.

The only remaining defence is the Intentional Fallacy from literary theory (Wimsatt and Beardsley, 1946): what matters in a work is not what the creator meant, but what the work contains, and what the viewer (or reader, hearer, etc.) experiences. In this context, it does not matter that the "aesthetic fitness function" of level 3 is meaningless (in our specific sense): no feeling/meaning is created, but none is needed. So then the "artistic system" is generating ideas and arbitrarily filtering them.

But what does this mean? An arbitrary, feeling/-meaningless "aethetic" function selects an arbitrary subset of the items that Colton's system would generate. Written another way: take the items that the system would generate and choose a random subset according to an arbitrary distibution. This is no different from generating arbitrary items. Thus, Colton's "artistic system" is doing nothing more than "mere generation" (Ventura, 2016), the most basic form of computational creativity, if one accepts it as computational creativity at all. Therefore, without a meaningful account of the aesthetic function, the distinct levels of Colton's hierarchy collapse into a single layer.

Alternatively, in Ventura's terms, Colton's level 2 is somewhere near "Algorithm 8 (…random generation … and filtering …)"[8] and "Algorithm 9 (…choosing a theme, … acceptable semantics …)". But because of the collapse of the "aesthetic fitness function", above, Colton's level 3 regresses, in Ventura's more precisely elaborated scale, to "Algorithm 4 (Generation …)"—to be explicit: generation *without* filtering, a step definitively backwards on Ventura's scale.

We've been Framed

Level 4 of Colton's roadmap is entitled "Persuasive Systems". Here, the designer has built "a *persuasive system* that can change your mind through explanations as well as high quality, surprising output." In this case, the arbitrary, even random, outputs of level 3 have influenced the viewer, and changed her (human) aesthetic sensibilities. A module is added for the software to generate explanations, so that the machine can explain what it did. It cannot explain in terms of feeling/meanings, because it has none, not even ersatz copies of human ones. So either it must explain in terms of syntactic generative steps (for there is nothing else), or it must *pretend* to have an aesthetic. The latter can easily be achieved by writing in words that relate to human aesthetics: "I felt…", "It seemed…", and so on. Colton's own system *The Painting Fool* (eg., Colton, 2012; Colton et al., 2015) and DARCI (Norton, Heath, and Ventura, 2013) both do this kind of text generation. DARCI has explicitly learned its aesthetic and its descriptive vocabulary from humans, and is thus emulating human feeling/meaning. It is less clear, at least to this author, what is the position with *The Painting Fool*; however, Colton's text suggests that the utterances are programmed, not learned, which gives them an ersatz feel. *The Painting Fool*'s website[9] contains extensive first-person writing, from the perspective of the system—but this text is written by a human, and not by the system.

In the context of his alternative question, Turing (1950, p. 434) considers the possibility of making a computer look more human to help it win the Imitation Game. He concludes that there would be "little point in trying to make a 'thinking machine' more human by dressing it up in … artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from seeing or touching the other competitors, or hearing their voices." So Turing concluded that, in order to answer the question, "Can a machine be mistaken for a human in a sustained written conversational competition with a human?", one should not frame the machine in a way that assisted its impersonation. Rather, one should be scientifically neutral, and prune away such confounding foliage.

Colton (2019, §3) presents a brilliantly effective explanation, entitled "Computational Authenticity", of how framing may change the perceived meaning of a poem. The poem, *Childbirth*, was generated by a computer. But Colton demonstrates how its meaning changes, depending on how it is framed. Its fictional author seems initially female, but then we are told the given name is a pseudonym for a man, and a criminal at that, and finally that neither author really exists. The demonstration is indeed powerful. But then, Colton explains, "We see fairly quickly that it is no longer possible to project feelings, background and experiences onto the author, and the poem has lost some of its value" (Colton, 2019, §3), and we see that he has in fact fallen into the Intentional Fallacy, and not wielded it as defence. Specifically: while the reader may well infer meaning in the poem from their knowledge (correct or otherwise) of its author, it is not what the author thinks that is important in the poem, but what the reader thinks. While projecting onto the maker of an artefact is indeed a pass-time that many humans relish, the resulting conclusions, correct or otherwise, are not part of the poem, but part of the viewer. Thus, they figure in an *external* evaluation, but not in an *internal* one, in respect of the creative system that produced the poem. The Romanticist notion that the "value" of a poem lies in projecting back on to what the author meant or in what they were thinking, was prevalent in the 19th century, but has not been so for more than 50 years (Wimsatt and Beardsley, 1946).

Consider the following thought experiment. The music of Pérotin, a member of the Notre Dame school of composition, around the turn of the 13th Century, is among the earliest surviving attributed music in the West. Almost nothing is

[7] The use of random and arbitrary choices in art is, of course, a valid aesthetic decision (Cage, 1973; Revill, 1993). But if that is the case here, then the decision is made not by the machine, but the programmer, contradicting Colton's premise.

[8] The ellipses in this paragraph hide parts of Ventura's definitions that are not specified in Colton's roadmap.

[9] `http://www.thepaintingfool.com`

known of this person—even his birthdate and nationality are uncertain. Has Perotin's music "lost some of its value" because we have no information about him on which to base our own interpretation? Apparently not: his music survives, and is still performed and recorded, after 700 years, a truly exceptional duration in Western culture. Of course, one might argue that the very lack of information contributes value, or at least mystique. But that is the exact opposite of the argument that Colton (2019) is explicitly and unambiguously proposing, so does not refute the Perotin counter-argument.

Digging deeper into Colton's argument about *Childbirth*, one sees a pattern. Initially, the poem is presented as the description of life experience by a woman, entering motherhood (deemed, along with apple pie, as "always good"). We are shown what appears to be an expression of feeling about something wonderful, and on which we all vitally depend: we form our own internal explanation of this meaning, as soon as it is offered. Thus, the affective response invoked is not only invoked by the poem, but by the intensification of our emotional connection with motherhood—which is peripheral to the poem. Next, the mother is violently torn away from us and replaced by a repellant person, and we are told that the poem is now about his repellant acts. The poem has not changed, but it is now associated with an explanation that most people will find unpleasant, and that unpleasantness is amplified by contrast with, and loss of, a feeling/meaning of noble and beautiful motherhood. Rhetorical success is clearly afoot, but that success is directly due to Colton, and not at all to his program. Furthermore, the relief that we feel when we learn that the poem was in reality constructed by a machine, and not by a repellant criminal, becomes the central affect, eclipsing the more important fact that really quite a good poem has been produced by a simple computational "cut-up" technique.

The problem here is that the framing of the poem, and the demonstration of its change, while vivid and cleverly executed, is functioning like political "deadcatting"[10], (mis)directing our attention away from the important point: the feeling/meanings that really are generated in each individual who reads the poem. That the poem was produced by a cut-up technique is surprising: most such poems will be (much) less good than this one, by chance, so the likelihood is that the outputs of this system were curated, leading us back, again, to Ventura's pre-creative Algorithm 4. With the dead cat of imaginary authors, Colton directs our attention away from the really interesting possibility: a computational system, capable of representing and reasoning about the syntactic and semantic patterns, and other more abstract images, that are suggested by chance in this poem, and then *selecting* this poem from other random outputs of the same random process as something of value. That would be Ventura Algorithm 8 or beyond.

Colton and colleagues suggest that creative machines making artefacts about human-centric issues will "naturally be seen as inauthentic" (Colton et al., 2020), in a classic and extreme application of the Intentional Fallacy. To refute this: a further thought experiment has a different man writing *Childbirth*. This man, aged 60, is a celibate, cloistered, Trappist monk, with no experience of women, nor of the outside world since age 16, and, therefore, no direct experience of childbirth or any of the associated social mores. If this man had written the poem, would it be "inauthentic"?

Of course not. If we frame the poem with knowledge of the monk, we can see it as a vision of a different life, that he never experienced, or even a religious expression, which is deeply felt and believed, but, in the cold light of day, still not experienced. The construction of such an image in the mind of the monk is *no less abstract* than the symbols used to infer a corresponding structure in a computer, despite the fact that he probably experiences feeling/meaning as a result of them, while the computer does not.

This thought experiment demonstrates that a poet's lack of experience of a thing does not render their poem about that thing inauthentic. Indeed, that lack of experience could, for some people, make the poem *more* remarkable. This applies as much in scientific creativity as in poetic creativity: Einstein did not have the opportunity to experience his physics directly, but imagined abstract things, initially internally, through thought experiments, then externally through mathematics. Only after his death were his ideas empirically validated. Einstein's abstraction did not make his ideas inauthentic; on the contrary, it made them all the more amazing.

This ersatz notion of "authenticity", which we suggest is misguided, leads to even more moot philosophy, relating to consciousness, intelligence and humans' relationships to computers. Authenticity is indeed important to humans, since it is related to trust, and thence comes the current interest in Explainable AI. What makes things authentic, to most humans, is truth—not artistic or absolutist notions of truth, but simply a thing being what it claims to be. If a man is a known liar, his authenticity is doubted by others, and they do not trust him. If a product does not do what it says on the tin, it does not sell for long. A human artist may construct a persona for herself, and present her art in that context, but if there is a lack of truth in that persona, then the artist is likely to lose the trust of her audience[11]. This human construct of trust is relevant to the idea of framing in computational creativity, and to artificial intelligence in general.

A Pig in Lipstick

"Framing" is an ambiguous word. It can mean "explicitly placing a created artefact or concept in a particular context". It can also mean "diverting the suspicion of a crime on to another person", another kind of misdirection. Misdirection of this kind, if exposed, will backfire on the perpetrator.

Colton et al. (2020) propose that computational creativity should adopt the idea of The Machine Condition, by analogy with The Human Condition, with the laudable aim of helping people to relate better to computers. The essential idea is to make computers seem more human-like by attributing their actions to their "life experiences". Colton et al. (2020) assert

[10] `https://en.wikipedia.org/wiki/Dead_cat_strategy`

[11] Lack of detail of such a persona seems more effective than detail, because fans may project what they like on to it. But, then, it is easier to lose faith in a projection than in a reality, when reality intrudes.

that "an entity like a machine does not need to satisfy notions of being alive or conscious to have life experiences worthy of communication through creative expression." For this not to be an absurd contradiction in terms, we must take "life experience" as a metaphor—for otherwise, how can something that is neither alive nor experiencing consciousness have one? So this is an ersatz notion of life experience, accompanied by no feeling/meaning. Ultimately, the problem is that consciousness is a defining prerequisite of the human condition: what it is like to be a human. There is no machine condition, *unless it be a fake one*, because it is not like anything to be a machine.

In summary, the idea is to *computationally create* framing like that of *Childbirth* (Colton, 2019) as the background to artificial intelligence. The framing would be computationally created, but as artificial as the intelligence that the machine may exhibit. While the facts on which the framing is based may be true, there is no sense in which they or inferences from them are true life experiences, any more than the monk has true life experiences of childbirth, above. If the monk presented these as true life experiences, we would call him a liar, and lose trust in him. The computer has experienced less than the monk—indeed, nothing at all.

Note the important difference between this framing, and the framing of artistic and scientific work in the human world. While human creators may indeed write about their own work, their writing is explicitly presented as such, and not as, for example, authoritative programme notes or exhibition guides. True framing comes from *outside* the creative system; it arises not from the actions of the creative system itself, but from the social milieu in which the creativity is taking place. If a human artist presented his own writing as the programme note of a critic, he would be a fraud: therefore, framing of this special kind, arising from the creator itself must be explicitly signalled as such, if it is to be honestly presented.

If appropriately and carefully presented, framing can be helpful in understanding. It can also be an entertaining fiction. If left to stand unexplained, or improperly attributed, it is fraudulent and also fundamentally misleading. The danger is only magnified by the beguiling effect of human anthropomorphic illusions about computing machinery. There is little enough understanding of the true nature of computers in the general population, without obfuscating it by pretending, or, worse, faking, humanity and consciousness. Using these terms, even while acknowledging their untruth, is both logically unsound and deceitful at the same time.

Furthermore, a pig in lipstick remains a pig. Eventually, however florid and beautifully gilded the frame, people will see the untruth of the picture it surrounds. Computational Creativity, as a research field, will suffer greatly if its human audience comes to believe it is fraudulent, and more so if the misdirection is deliberate. The frame, even if computationally created, must not obscure even the edges of the truth.

Consequences

What are the consequences of these arguments for Computational Creativity? Some desiderata are now proposed.

1. Creativity is not an *ex ante* phenomenon, and our research field knows this. We seek what is *necessary in general* for the perception of creativity, not what is *sufficient to a particular case*. Let us make this clear.
2. The fact that creativity is an *ex post* phenomenon, involving the perception of humans, does not entail that we should focus on *manipulating* that perception. On the contrary, let us investigate the necessary properties for creativity thoroughly, and test them openly and honestly, without obfuscation—even obfuscation that is computationally created.
3. Creativity, intelligence, and consciousness are inextricably linked in humans. While accepting the lack of consciousness in computers, let us study the relationship between creativity and intelligence in the light of that knowledge, with rigorous, philosophically careful arguments, such as those of Turing, seventy years ago.
4. Embracing human-based aesthetics does not prevent a computational system from surprising us or changing our personal aesthetic; indeed, knowing about human-based aesthetics is the first step to reliably challenging their *status quo*. Since computers are not capable of feeling/meaning, let us not be shy of human-based aesthetics, for we have no alternatives. Let us instead challenge humanity on its own aesthetic terms.
5. Constructing computational aesthetic measures, to be used in a creative context, and managing their interaction with AI techniques that we use for our creative systems, is non-trivial. Let us not be beguiled into framing the shortcomings out of our systems. Let us focus our limited resources, not on illusions, but on solutions.
6. Let us remember that a pig in lipstick remains a pig.

Ultimately, framing based on pretence and philosophical falsehood, no matter how well executed, no matter how well-intentioned, is a beguiling Yellow-Brick Road to an Emerald City of creative systems. At some point, someone will pull back the curtain, the Wizard will be exposed as a fake, and the story will end.

Epilogue and Challenge

On reworking this paper following helpful reviews, I understood, to my surprise, that a key motivation for misleading framing is in fact the *ex post* nature of creativity itself. If we seek systems that humans will "deem to be creative" (Colton and Wiggins, 2012), then there is always the option of convincing those judges by sleight of hand, supplying what is *sufficient* for the perception of creativity in a given case, instead of focusing on the *necessary* computational components and processes that will enable scientific progress towards creative behaviour in computers.

Inappropriate framing of this kind is unlikely to succeed enduringly, because such sufficient properties are, I think, likely to be context-dependent and so unlikely to be general. Therefore the illusion will quickly fade. But this impermanence only renders the clear danger of discovery more present, placing trust in our research field at yet greater risk. It seems, then, that, in order to define our field correctly, we need a better way of acknowledging its the *ex post* nature, so as to address both my alternative interpretation of the Imitation Game, and the problem of inappropriate framing in Computational Creativity.

Acknowledgments

The ideas presented here were influenced by insightful conversations with Alan Smaill, at the University of Edinburgh, Graham White, at Queen Mary University of London, and Anna Kantosalo, Simo Linkola and Hannu Toivonen at the University of Helsinki, though I do not claim their agreement with the content of this paper. I am very grateful for these wise and well-informed discussions, and for the remarks of three anonymous reviewers, whose suggestions clearly improved the paper.

References

Bellman, R. E. 1978. *An Introduction to Artificial Intelligence: Can Computers Think?* San Francisco: Boyd & Fraser Publishing Company.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London, UK: Routledge, 2nd edition.

Cage, J. 1973. *Silence*. Wesleyan University Press.

Charniak, E., and McDermott, D. 1985. *Introduction to Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? In de Raedt, L.; Bessiere, C.; Dubois, D.; and Doherty, P., eds., *Proceedings of ECAI Frontiers*.

Colton, S.; Halskov, J.; Ventura, D.; Gouldstone, I.; Cook, M.; and Ferrer, B. 2015. The painting fool sees! new projects with the automated painter. In *ICCC*.

Colton, S.; Pease, A.; Guckelsberger, C.; McCormack, J.; and Llano, M. T. 2020. On the machine condition and its creative expression. In *Proceedings of the International Conference on Computational Creativity*, 342–349.

Colton, S. 2012. The painting fool: Stories from building an automated artist. In McCormack, J., and d'Inverno, M., eds., *Computers and Creativity*. Springer-Verlag.

Colton, S. 2019. From computational creativity to creative ai and back again. *Interalia Magazine*.

Countess of Lovelace, A. 1842. Translator's notes to an article on babbage's analytical engine. *Scientific Memoirs* 3:691–731.

Dennett, D. 1991. *Consciousness explained*. Boston: Little, Brown and Co.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hodges, A. 1992. *Alan Turing: The Enigma*. London: Vintage.

Hodson, J. 2017. The creative machine. In *Proceedings of the International Conference on Computational Creativity*. Association for Computational Creativity.

Hoenig, F. 2005. Defining computational aesthetics. In Neumann, L.; Sbert, M.; Gooch, B.; and Purgathofer, W., eds., *Computational Aesthetics in Graphics, Visualization and Imaging*. EuroGraphics.

Jefferson, G. 1949. The mind of mechanical man. *British Medical Journal* 1:1105–1121. Lister Oration for 1949.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4(3):246–279.

Jordanous, A. 2018. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the AISB 2018 Symposium on Computational Creativity*.

Kantosalo, A., and Toivonen, H. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the Seventh International Conference on Computational Creativity*.

Merker, B.; Williford, K.; and Rudrauf, D. 2021. The integrated information theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences* 1–72.

Merker, B. 2007. Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences* 30(1):63–81.

Merker, B. 2013. The efference cascade, consciousness, and its self: naturalizing the first person pivot of action control. *Frontiers in Psychology* 4(501).

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

Nagel, T. 1986. *The View From Nowhere*. Oxford University Press.

Norton, D.; Heath, D.; and Ventura, D. 2013. Finding creativity in an artificial artist. *Journal of Creative Behavior* 47(2):106–124.

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

Preston, J., and Bishop, M. 2002. *Views into the Chinese Room*. Oxford: Oxford University Press.

Revill, D. 1993. *The Roaring Silence: John Cage: A Life*. Arcade Publishing.

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

Russell, S., and Norvig, P. 1995. *Artificial Intelligence – a modern approach*. New Jersey: Prentice Hall.

Searle, J. 1980. Minds, brains and programs. *Behavioral and Brain Sciences* 3(3):417–457.

Searle, J. 1999. *Mind, language and society*. New York, NY: Basic Books.

Shanahan, M. 2010. *Embodiment and the inner life: Cognition and Consciousness in the space of possible minds*. OUP.

Tononi, G. 2004. An information integration theory of consciousness. *BMC Neuroscience* 5(42).

Turing, A. 1950. Computing machinery and intelligence. *Mind* LIX(236):433–60.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the International Conference on Computational Creativity*, 17–24.

Wiggins, G. A.; Tyack, P.; Scharff, C.; and Rohrmeier, M. 2015. The evolutionary roots of creativity: mechanisms and motivations. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370(1664).

Wiggins, G. A. 2006a. A preliminary framework for description, analysis and comparison of creative systems. *Journal of Knowledge-Based Systems* 19(7):449–458.

Wiggins, G. A. 2006b. Searching for computational creativity. *New Generation Computing* 24(3):209–222.

Wimsatt, W. K., and Beardsley, M. C. 1946. The intentional fallacy. *The Sewanee Review* 54(3):468–488.

Winston, P. H. 1992. *Artificial Intelligence*. Reading, Massachusetts: Addison-Wesley, 3rd edition edition.

Wittgenstein, L. 1958. *Philosophical Investigations*. Oxford, UK: Blackwell.