Evaluating Computational Creativity: An Interdisciplinary Tutorial

CAROLYN LAMB, DANIEL G. BROWN, and CHARLES L. A. CLARKE, University of Waterloo

This article is a tutorial for researchers who are designing software to perform a creative task and want to evaluate their system using interdisciplinary theories of creativity. Researchers who study human creativity have a great deal to offer computational creativity. We summarize perspectives from psychology, philosophy, cognitive science, and computer science as to how creativity can be measured both in humans and in computers. We survey how these perspectives have been used in computational creativity research and make recommendations for how they should be used.

Categories and Subject Descriptors: J.5 [**Applied Computing**]: Arts and Humanities

General Terms: Human factors, Measurement, Theory

Additional Key Words and Phrases: Computational aesthetics, computational creativity, digital art

ACM Reference format:
Carolyn Lamb, Daniel G. Brown, and Charles L. A. Clarke. 2018. Evaluating Computational Creativity: An Interdisciplinary Tutorial. *ACM Comput. Surv.* 51, 2, Article 28 (February 2018), 34 pages.
https://doi.org/10.1145/3167476

1 INTRODUCTION

Computational creativity is *the philosophy, science and engineering of computational systems which, by taking on particular responsibilities, exhibit behaviors that unbiased observers would deem to be creative* (Colton and Wiggins 2012). Computational creativity is a growing field, and creative systems have been developed for applications ranging from music, visual art, and poetry to mathematics, design, and code generation (Loughran and ONeill 2017).

However, many attempts at computational creativity lack rigor, especially in evaluation. A major contributor to this situation is a lack of interdisciplinary knowledge about creativity. Psychologists, philosophers, cognitive scientists, and others have attempted to tackle the problem of creativity evaluation from their own perspectives, and computer scientists working on creative systems can learn from their efforts.

As yet, there is no consensus on how to evaluate a creative system. Many interesting theoretical proposals are available (Aguilar and Pérez y Pérez 2014; Bown 2014; Burns 2015; Colton 2008; Gervás 2002; Jordanous 2012a; Negrete-Yankelevich and Morales-Zaragoza 2014; Pease et al. 2001;

Authors' addresses: C. Lamb, D. G. Brown, and C. L. A. Clarke, David Cheriton School of Computer Science, University of Waterloo, 200 University Avenue, Waterloo, Ontario N2L 3G1; emails: carolyn.elizabeth.lamb@gmail.com, dan.brown@uwaterloo.ca, claclark@gmail.com.

ACM 0360-0300/2018/02-ART28 $15.00
https://doi.org/10.1145/3167476

Ritchie 2001), but the reliability and validity of many of these proposals are in question. Many are not well grounded in the psychology or philosophy of human creativity. (There is, of course, an argument that computational creativity need not resemble human creativity—an argument which we will explore in Section 7—but it is inadvisable to reject theories of human creativity without first understanding them.)

This proliferation of potentially unreliable evaluations can leave a researcher at a loss for where to begin. It is tempting to rely on an *ad hoc* method. Further, a computer scientist may be daunted by the breadth and depth of research to assimilate to form one's own theories.

Computer scientists may not have the time or inclination to exhaustively study the psychology of creativity, so we thoroughly summarize relevant research which can be used as a guide. We approach evaluation ideas from an interdisciplinary perspective–from psychology, philosophy, and other fields, as well as from the computational creativity research community itself. After reading this article, a researcher should understand the different ways a human or computer might evaluate the presence of creativity, and where to look for more information. This should make it easier to form informed opinions about computational creativity and decide what kind of evaluation is appropriate for one's project.

In Sections 2–6, we introduce theories of creativity evaluation from four perpectives—including both human contexts and computational ones. We follow this, in Section 7, with a discussion of questions as to whether creativity can be evaluated at all. Finally, in Section 8, we discuss practical issues specific to computational creativity and some common pitfalls of evaluation in practice. Each section has specific takeaways for the researcher who wants to apply the ideas to their work. These takeaways will be summarized in our conclusion, Section 9.

2 THEORIES OF CREATIVITY

A variety of theories of creativity evaluation exist. When comparing evaluations to each other, it is useful to group them based on their theoretical perspective. One useful taxonomy of theories, which we use to structure this article, is known as the four P's: Person, Process, Product, and Press (Rhodes 1961).

- **Person** is the human (or non-human agent) who is seen as creative. Person theories study what it is about the agent that makes them creative.
- **Process** is the set of internal and external actions the agent takes when producing a creative artifact. Process theories study what sort of actions are undertaken when creative work is done.
- **Product** is an artifact, such as an artwork or a mathematical theorem, which is seen as creative or as having been produced by creativity. Product theories study what it is about the product that makes it worthy of being called creative.
- **Press** is the surrounding culture which influences people, processes, and products and which judges them as creative or uncreative. Press theories study what it is that leads a culture to view something as creative.

These four P's originate with Rhodes (1961) and were introduced to computational creativity by Jordanous (2016a). Jordanous suggests the use of the word "producer" in place of "person," to emphasize that the creative agent need not be a human. For this article, we use "person" so as to match the psychological literature.

We will describe evaluation methods from each of these perspectives in turn. Of course, it is possible to evaluate from more than one perspective. For example, Colton et al. (2011) introduce, in the same paper, both the FACE (process) model for assessing the creativity of an act, and the IDEA (press) model for assessing its impact. The SPECS model (Jordanous 2012a), as we will discuss, contains criteria that arguably cover all four perspectives (Jordanous 2016c). Many evaluations that are mainly in one perspective also incorporate ideas from others. However, in order to be clear about the ideas behind each perspective, we will for the most part discuss them separately.

3 PERSON PERSPECTIVE

The Person perspective is grounded in psychometrics (the measurement of human mental traits). The goal of the Person approach is to discover what personality, emotional, and cognitive traits distinguish a more creative person from a less creative one. The Person perspective is potentially useful for researchers who want their systems to be viewed as creative agents because of their inherent traits.

For humans, hundreds of psychometric tests for creativity exist (Plucker and Renzulli 1999), the most famous perhaps being the Torrance Tests (Torrance 1968). These tests prompt the person to generate many related ideas—for example, by asking how many ways one can use a chair. This constitutes a test of *divergent thinking*—the ability to come up with varied and unusual ideas—which is a popular Person-based definition of creativity (Cropley and Cropley 2005). However, in many theories, divergent thinking is only useful when it alternates with convergent thinking—the use of conventional knowledge to evaluate and develop the ideas (Csikszentmihalyi 1996).

Computers can pass some psychometric creativity tests. Olteţeanu and Falomir's system comRAT-C solves a Remote Associates Test by using associative spreading through a database of common bigrams (Olteţeanu and Falomir 2015). Their system OROC passes an Alternative Uses Test by comparing information about objects' uses and physical properties (Olteţeanu and Falomir 2016). Gross et al. (2012) solve a Remote Associates Test using mined word associations, and discuss how this technique could be used in other creative tasks. Psychometric tests more distantly related to creativity, such as the analogical reasoning of Raven's Progressive Matrices, have also been the subject of computational creativity research (Johner et al. 2015). However, these systems do not do other creative work besides taking tests. Since the structure of a computer system differs from the structure of a human mind, it is not evident that the success of a computer at a psychometric test, *per se*, is a useful proxy for its general creative ability—nor are the above authors making that claim.

Another avenue of Person research is to study famous creative people. This is known as the historiometric approach. Historiometric research reveals multiple successful approaches to creativity (Policastro and Gardner 1999), with a few general traits in common, such as an ability to distinguish promising avenues of work from less promising ones.

Policastro and Gardner (1999) review historiometric studies and sort creators into four categories, based on two distinctions: creators who work with objects and symbols versus those who work with people, and those who strive for excellence within a domain versus those who challenge the domain's foundations. Each combination of these traits produces a creative personality type. Masters, who achieve excellence within an object- or symbol-based domain, include Mozart, Rembrandt, and Shakespeare. Makers, who create a new object- or symbol-based domain, include scientists like Einstein and Darwin as well as maverick artists like Stravinsky or Joyce. Introspectors, who work for excellence in a people-based domain, use art to express their inner selves; these include Proust and Woolf. Their counterparts, the Influencers, work to change and challenge other people. Influencers include Gandhi, Mandela, and Eleanor Roosevelt. In Policastro and Gardner's theory, those who challenge a domain's foundations are not necessarily more creative than others, only creative in a different way. This is a point that we will return to in Section 5.1.1.

Measuring the personal traits of computers is conceptually difficult, but some researchers have attempted it. The Creative Tripod (Colton 2008) works by attributing traits to a system, but since the Tripod's focus is on convincing people that the system has these characteristics, it will be discussed under Press in Section 6.1. A few researchers have attempted to model traits associated with creativity, such as curiosity (Grace et al. 2017). More commonly, as with the Tripod, attributing traits to a computer is something humans do as the result of information from one of the other three perspectives.

Takeaway: If your system is meant to exhibit specific human cognitive traits, use existing tests to measure those traits. Otherwise, claims about your system should likely be based on another perspective.

4 PROCESS PERSPECTIVE

The Process perspective includes any theory of *how* creative products are made—that is, what cognitive steps must be taken in order for an activity to be creative. Many of these theories are descriptive rather than evaluative. Since they are created by researchers studying humans, some are built on the assumption that a human cognitive structure, including unconscious reasoning, is already in place. Thus, these ideas have not always been straightforwardly taken up in computational creativity. However, the Process perspective includes many important ideas that can and should influence the design of a creative system. The Process perspective is especially useful for researchers who want to model human creativity, or who want to make an argument that their system is creative because of the kinds of tasks it does.

4.1 Conceptual Space

The most popular Process theory in computational creativity is Boden's (1990) theory of conceptual space. Boden divides creativity into three types: combinatorial, exploratory, and transformational. Combinatorial creativity occurs when two familiar ideas are put together in an unfamiliar way. The remaining types of creativity depend on the idea of a conceptual space: the space of all things that could be generated according to a set of rules. Exploratory creativity, by testing out the rules' implications, reaches accessible points in the space that have not been reached before. Transformational creativity changes the rules to reach points that were not accessible before. Wiggins (2006) refines Boden's idea of conceptual space and describes types of "aberration" which might lead one to revise the space and/or the way of searching.

Many Process theorists privilege transformational creativity. Several philosophers, for example, define originality as work that changes established rules (Bartel 1985). To these philosophers, transformational creativity is more original than exploratory creativity. Dorin and Korb (2012) define creativity as "the introduction and use of a framework that has a relatively high probability of producing representations of patterns that can arise only with a smaller probability in previously existing frameworks." The more unlikely the new patterns were under previous frameworks, the more creative they are. Alternatively, a number of computational creativity researchers focus on combinatorial creativity, teaching a computer to blend seemingly unrelated concepts (Cunha et al. 2017; Gonçalves et al. 2017; Veale 2013).

Besold (2016) argues that certain processes in machine learning, such as Bayesian theory learning and inductive logic programming, are transformationally creative, as they involve the generation of novel (to the system) ideas. Besold argues that computational creativity, particularly scientific and problem-solving creativity, should use these processes.

Takeaway: Think about how your system explores its conceptual space. What are the parameters of the space? Do you want it to combine existing ideas, to generate something radically transformative, or to explore the space more modestly?

4.2 Stage- and Loop-based Theories

The theories described above are vague as to how a human goes about transforming a conceptual space. However, more naturalistic models of the creative process are plentiful. One such theory is

Wallas' (1926) four-stage theory: preparation, incubation, inspiration, and verification. The four-stage theory is based on case studies of how scientists have ideas. First, in preparation, the creative person gathers information related to their task. In incubation, the person ponders the information, makes connections, and often abandons the project for a time, while further connections are made unconsciously. In inspiration, something "clicks," and the person gets an idea based on a novel way of looking at the information. Then in verification, the person does a lot of hard work establishing, developing, and polishing their idea into a finished product.

Sadler-Smith (2015) suggests that the case studies on which the four-stage model is based also include a fifth stage, between incubation and inspiration: intimation. At the intimation stage, a solution to the problem exists in fringe consciousness, but has not yet been seized on by the conscious mind. There may be a feeling that a solution is coming, or an awareness of parts of the solution at the edge of the mind. According to Sadler-Smith, expertise helps a creative person progress to the inspiration stage by unconsciously evaluating their unconscious ideas. Csikzentmihalyi (1996) includes between inspiration and verification the stage of evaluation, in which the creative person uses their domain knowledge to decide if their idea is worth pursuing.

In practice, these elements do not necessarily follow each other in a tidy order, but are mixed together (Beardsley 1965) or can repeat "fractally" as a series of smaller and smaller inspirations about details of the work (Csikszentmihalyi 1996). This mixing leads researchers to consider creativity as a loop or an iterative process—but what characterizes the iteration?

One iterative theory is BVSR—Blind Variation, Selective Retention. BVSR was first proposed by Campbell (1960) and was further developed by Simonton (2011). In BVSR, a creative agent blindly generates many possible ideas before reflecting and choosing the best ones. For generation to be "blind" means that the probability of generating an outcome is decoupled from its utility: an increase in the utility of an outcome does not cause an increase in the probability of generating that outcome. This is also called the Darwinian theory of creativity (Kronfeldner 2010) because of its superficial resemblance to natural selection, in which random mutations and recombinations create organisms which are selected for their fitness. However, a close correspondence between BVSR and evolution is not strictly necessary (Simonton 2011).

Other psychologists and philosophers (e.g., Dasgupta (2011) and Kronfeldner (2010)) have criticized BVSR. The main criticism involves the notion that generation is blind. Dasgupta (2011) outlines famous cases of creativity where candidate solutions were generated based on schemas, building on partially successful previous ideas to form a progressively more correct solution. These mechanisms are not explained by BVSR. Mechanisms like analogical reasoning and spreading activation, which Simonton lists as examples of blind processes, are not actually blind: they depend on the structure of knowledge in an agent's mind, and on the thinking the agent has already done about the problem, which is likely to guide them toward more useful solutions (Kronfeldner 2010).

Despite these problems with the concept of blindness, most theories of the creative process still involve a loop between generation and evaluation. (Note that the alternation between generation and evaluation is roughly equivalent to the alternation between divergent and convergent thinking mentioned in Section 3.)

Two cases in point are the Geneplore model (Ward et al. 1999) and the ER model (García et al. 2006). In Geneplore—a portmanteau of "generate" and "explore"—an agent generates preinventive structures through some means, not necessarily blind: synthesis, transformation, and exemplar retrieval are mentioned (Ward et al. 1999). The agent then evaluates the preinventive structures and explores their properties and implications. This exploration leads either to refining the preinventive structures or discarding them and generating new ones. In the ER model—standing for Engagement-Reflection—the agent brainstorms ideas for solving a problem in the Engagement stage, then evaluates them at the Reflection stage (García et al. 2006). Ideas at the Engagement stage are not necessarily blind, but they are not evaluated at this stage. Ideas at the Reflection stage are discarded if their prerequisites are not met, or modified to make them more acceptable. They then form a partial solution to the problem. This partial solution is fed back into Engagement for elaboration, which is again reflected on, and so on, until the solution is complete.

Dahlstedt (2012), similarly, defines the two phases as "implementation" and "re-conceptualization." During implementation, an artist has two ideas: a conceptual description of the desired end product, and an instruction for how to get there. In some genres, such as improvisatory theater, the description may be missing. In re-conceptualization, the artist compares their work to their conceptual description, and changes either the product or the description to bring them closer to each other. These changes can take various forms, such as additions, expansions, generalizations, mutations, new constraints, replacements of material, or even a new conceptual representation from scratch. When the work is finished, the conceptual representation is hidden; each person in the audience then makes their own conceptual representation of what they think the art is about.

The cultural psychologist Glăveanu (2015) defines these loops as being based in perspective-taking: an artist must switch between the perspective of a creator and the perspective of an audience viewing the art. An artist has interactions with real audience members throughout their career, which allows them to broaden the range of perspectives they can take on when evaluating their work. Basing the creative loop on perspective-taking defines it as a social phenomenon which is inextricably connected to the Press perspective. A very simple form of perspective-taking takes place in systems which evaluate their work based on data about what humans prefer. More advanced social reasoning would take place in a system that can actively interact with critics; we describe a few such systems in Section 6.5.

Amabile's (2012) Componential Theory breaks the creative process into five steps: Task Identification, Preparation, Response Generation, Response Validation, and Communication. Some or all steps can be repeated if more progress is needed, and each step is affected by the personality and environment of the creator.

Systems based on these loops are common. A genetic algorithm naturally switches between generation of artifacts and evaluation of the current population, and these algorithms are popular (Galanter 2012). The ER model has been used to generate stories and do geometry problems. A few researchers (Gervás 2013) incorporate more critical evaluation, attempting to detect specific mistakes in their work and correct them. This type of evaluation, while difficult to implement, is more sophisticated than the random recombination of a genetic algorithm, and we recommend its use when possible.

We are unaware of any systems that explicitly use Wallas' (1926), Sadler's (2015), or Amabile's (2012) multi-stage models. The results of a study replicating these methods in a computer system might be enlightening.

Takeaway: Consider building your system based on a generation-evaluation loop or a more complex looping process. The evaluation stage in the loop should lead it to incrementally improve.

4.3 The Process of Professional Artists

These theories are supplemented by naturalistic studies of artists at work. Fayena-Tawil et al. (2011) summarize studies of artists and non-artists asked to create a drawing in the laboratory. (The creativity of the artists in this and similar studies is not measured; rather, it is assumed that working artists are creative.) Compared to non-artists, artists spend more time examining, selecting, and rejecting available objects; reworking their drawing; developing an overall composition; stating large-scale goals; and making large-scale evaluations. Non-artists spend most of their time trying to reproduce visual details.

Mace and Ward (2002) study self-reports by artists making art for a gallery. They divide the process into four stages: Conception, Development, Creation, and Finishing. In Conception, the artist gets an idea, either spontaneously, or by expanding on previous ideas. In Development, the artist does preliminary sketches, and enriches the concept through further associations, until they have a specific sense of what the work will look like. They then, in Creation, physically construct the artwork itself. This involves a generate-evaluate loop, with frequent restructuring in response to mistakes or new ideas. Finally, in Finishing, the artist prepares the art for display by framing, mounting, and so forth. Each stage includes the possibility of shelving the idea or returning to earlier stages.

The process may be different for different domains. Bourgeois-Bougrine et al. (2014) study the process of French screenwriters. For these writers there are three stages: Impregnation, Structure/Planning, and Writing/Rewriting. A screenwriter's Impregnation differs from an artist's Conception because screenwriters do not typically have the ideas for their own films: instead, a director hires them to create a screenplay on a specific topic. The Impregnation phase also includes rest or seemingly unrelated tasks. During this time, the screenwriter does not appear to be creating anything, but is actually allowing the vital incubation stage to happen. Next, during Structure/Planning, the writer decides on a tentative structure for the film. Finally, the writer actually writes the screenplay. Scenes are constantly rewritten as the writer finishes other scenes and traces their implications. Rewriting goes on throughout the filming, so there is no Finishing stage.

In spite of their differences, one can see a parallel in the progression from idea to structure to specific creation in both studies. It appears to be typical for creative humans to progress in this way, beginning with a simple inspiration, then planning and implementing.

Takeaway: Consider building your system to move from inspiration to planning to creation rather than trying to generate a full artifact all at once.

4.4 Autonomy

A creative system need not be a slavish imitation of human creativity. Researchers may instead build a system whose process emphasizes the strengths of computers (Gervás 2010). But one must also address the weaknesses of computers. One issue here is autonomy: the ability of the machine to decide what to do for itself.

A lack of autonomy is a major criticism both of general AI and computational creativity specifically. Mumford and Ventura (2015), surveying public opinion about creative computers, found that autonomy was one of the biggest issues, and is a particular issue for skeptics. Current systems possess some autonomy—being able to add to their knowledge base, for instance—but cannot define or refine their own processes. Guckelsberger et al. (2017) describe a thought experiment asking a system why it made a creative decision—and then asking "why" again, recursively; all existing systems would eventually have to answer "because my programmer told me to."

Negrete-Yankelevich and Morales-Zaragoza (2014), in their Apprentice Framework, suggest moving a system through four stages, each with greater autonomy than the last: the toolkit (a set of processes to be used by human artists), the generator (a machine that can make finished or partially finished work), the apprentice (a generator which makes novel and/or valuable work at least sometimes, with some human curation), and the master (a generator which always makes finished, acceptable-to-experts work on its own). Colton (2012) describes this autonomy-based progression as "climbing the meta-mountain." One looks at the decisions humans make for the system and automates those; then looks at the decisions humans make for the revised system, and automates those; and so on, potentially ad infinitum. Colton et al. (2014) suggest drawing diagrams of a system's process so that judges can see what the system is responsible for, what choices it makes, and what is left to humans. Current diagrams could be compared to older ones to show an increase in autonomy over time.

Complete autonomy may not be readily achievable. Relevant to this discussion is Smithers (1997), who argues that the term “autonomy” in AI is misused. “Autonomy” in AI is used to describe mobility, lack of a direct controller, self-regulation in a narrow sense, or the ability to do certain things automatically. Smithers argues that this definition is too narrow: computer systems are not autonomous unless they possess self-lawmaking and self-identity. That is, autonomous systems are governed by rules that they create through interaction with their environment. Autonomy as self-lawmaking is congruent with the way the word “autonomy” is used in law, politics, medicine, and biology. If we accept Smithers’ argument, then creating an autonomous creative system would require climbing many levels of Colton’s meta-mountain indeed.

Some researchers have taken up this quest for high autonomy. Guckelsberger et al. (2017) define adaptive creativity, in which a truly creative agent must develop its goals on its own the way living organisms do. The agent must be embodied and have a precarious existence, which requires it to continuously interact with its environment, preserving itself by modifying itself or the conditions around it. Only values generated through these attempts at self-preservation, in Guckelsberger et al.’s view, can truly belong to the agent. The most successful agents exhibit behavior that is novel and valuable (see Section 5.1.4): to survive, they must respond flexibly to unexpected threats.

As a metric for evaluation, Guckelsberger et al.’s ideas have some weaknesses. As Guckelsberger et al. admit, the set of systems which are adaptively embodied is quite different from the set of systems which most human observers would intuitively deem creative. In fact, it is uncertain if human creativity would meet Guckelsberger et al.’s standards. Humans are embodied agents with a precarious existence whose creative predilections arise from natural selection pressures. But human creativity is social in nature, and often far removed from preserving viability in the moment. Given the culturally constructed nature of many human endeavors, it is unclear if a human would pass Guckelsberger et al.’s recursive “why?” test; most humans do at least some things, including creative things, because another human told them to. This does not mean that Guckelsberger et al.’s theory is not useful for low-level creativity in embodied agents. But it does rest on strict assumptions which may not be useful for every type of creativity. Researchers uninterested in building embodied agents can still achieve interesting things by using weaker autonomy requirements.

Takeaway: Decide on a level of autonomous decision making that it is realistic for your system to have at its current stage. Ensure that you understand which decisions are made by the computer and which by humans, and that your description of the system realistically represents this. Consider increasing your system’s autonomy in later versions.

4.5 Specific Evaluation Techniques

All of this Process research provides sound advice for system design, but not all of it is clearly applicable to evaluation. Comparing one’s system’s processes to theoretical processes is useful as a check by the researcher, but it is not a formal, falsifiable evaluation. Fortunately, several evaluation methods exist which are process-based, in that the evaluator takes into account the system’s inner workings.

One such method is the FACE model (Colton et al. 2011), which distinguishes different things that a system could create: concepts, expressions of concepts, aesthetic measures, framing information, or methods for generating any of these. Colton et al. suggest a number of ways FACE could be used to evaluate creativity, including a “cumulative way,” in which a system is more creative if it performs more types of generative acts, thereby taking control of more of its own decisions. Or, in the “comparative way,” some components are considered more creative than others. A third option is the “process way,” in which a system with a more creative process is deemed more creative. What these authors think of as a more creative process is not clearly defined, although in one example, they state that they prefer more autonomy. This lack of definition means the process way verges on a tautology. In fact, all the ways of evaluating using FACE are vague. Jordanous (2012b found that the FACE model ranked musical improvisation systems in the opposite order to other evaluation methods, and its categories are not necessarily relevant to what researchers want to achieve, which calls into question FACE's validity.

Jordanous's SPECS model (2012a) evaluates systems based on 14 factors identified through study of how humans define creativity. These include active involvement and persistence, dealing with uncertainty, domain competence, general intellect, generation of results, independence and freedom, intention and emotional involvement, originality, progression and development, social interaction and communication, spontaneity and subconscious processing, thinking and evaluation, value, and variety, divergence, and experimentation. Many are impossible to evaluate without knowledge of a program's inner workings. (Bhattacharjya (2016) divides the SPECS criteria among all four P's; Jordanous (2016c) says many criteria cross more than one perspective.)

The SPECS model leaves it up to researchers how they will use the criteria, but Jordanous provides a case study in which musicians, after some training in how to think about creativity, were given information about musical improvisation systems and examples of products the systems had produced and asked to rate each program on the 14 criteria. The criteria are not combined into a single creativity score, but instead provide a nuanced picture of what the system does and does not do well (Jordanous 2012b).

A more hierarchical model is Ventura's (2016), which categorizes creative systems into seven categories, each more creative than the last. Randomization (in which output is completely random) and Plagiarization (in which output is copied from existing artifacts) are the least creative. In Memorization, the system modifies existing artifacts, and in Generalization, it creates new artifacts based on rules that can be programmed in or discovered by the system. In Filtration, the system can evaluate its output with a fitness function; any system using a generation-evaluation loop is at least at the Filtration level. Even more advanced, in Inception, the system uses a knowledge base to inject deeper meaning into its artifacts. Finally, in Creation, the system has perceptual abilities and is able to create artifacts based on what it sees, hears, and experiences.

Ventura's model is not an evaluation technique, but a quick-and-dirty evaluation could be performed by asking an expert to place a system in one of the seven categories. The advantage of this model is that, being an ordinal scale of levels of creativity, it is very simple to use to argue that one system is more creative than another. The disadvantage is that it lacks nuance. If two systems are at the Filtration level, for instance, Ventura's model says nothing useful about the distinction between them. Ventura's model also cannot show that filtration (or any other process) is done well.

As we saw in Section 4.4, further process evaluations can be done by placing a system in the Apprentice Framework (Negrete-Yankelevich and Morales-Zaragoza 2014), or asking experts to analyze a diagram of the system's responsibilities (Colton et al. 2014).

Ventura's model and the FACE model are developed *a priori* from theory, and SPECS is based on a linguistic analysis of what humans associate with creativity. We do not know of an evaluation technique that comes at a process from a psychological perspective—that is, one that systematically compares a process to the human creative process described in Section 4.2. Although a creative computer need not exactly resemble a human, such a technique would still be an extremely interesting contribution.

Care must be taken to match the theory underlying a system to the theory underlying its evaluation. For instance, if one's goal is to make a system as autonomous as possible, one may not be interested in Ventura's model, which deals only indirectly with autonomy. The best use of process theories may not be a binary evaluation of a system as creative or not; most forms of process theory do not work that way. Instead, process-based evaluation forms a nuanced picture of what a system does and does not do for itself.

Takeaway: Process evaluations tend to be either a placement of the system in one of several broad categories, or a qualitative analysis of the system's process strengths and weaknesses.

5 PRODUCT PERSPECTIVE

Rather than Person or Process (or Press), many evaluations in computational creativity focus on Product. Even a system with a very sophisticated process must at some point produce a creative artifact. Intuitively, if a system produces artifacts that are never any good, it is hard to argue that the system is very creative. We can evaluate the product itself as creative or uncreative by assessing its traits.

Note that, by "artifact," we do not only mean artworks: a mathematical theorem, scientific hypothesis, business plan, or engineering design is an artifact just as poems, visual artworks, and pieces of music are. Even a way of adapting to the environment is arguably a creative artifact (Guckelsberger et al. 2017; Aguilar and Pérez y Pérez 2014). However, it is assumed in the Product perspective that all creative systems, in some way, produce something.

The Product perspective is especially useful for systems that have the goal of producing something useful to humans.

5.1 Novelty and Value

A very common set of creativity criteria are novelty and value. That is to say, if and only if a product is both novel and valuable, is it considered creative. This is the leading definition of creativity among philosophers and psychologists, and has a long history (Gaut 2010); it was introduced to computational creativity by Boden (1990). Jordanous (2016a) points out that novelty and value can be applied to each of the four P's: a system can employ a novel process, interact with the press in novel ways, or exhibit novel personal traits related to creativity. However, we will focus here on novel and valuable products. We will look at both concepts in depth before discussing how they have been brought together in practice.

5.1.1 Novelty. Novelty has several definitions. Boden (1990) distinguishes between P-creativity—that which is novel because its creator has not thought of it before—and H-creativity—that which is novel because no one has thought of it before. Boden also raises the issue of "mere novelty"—products which are novel, but in a trivial or uninteresting way. Boden suggests the use of *surprise* in addition to novelty: an idea is more surprising when it requires a more fundamental change to conceptual space. Note that this definition, and others, privilege transformational over exploratory creativity! Simonton (2011) similarly deals with mere novelty by citing U.S. patent law: a product is creative when it is novel, valuable, and non-obvious (based on pre-existing domain-specific knowledge).

Along with the distinction between P-creativity and H-creativity comes the finer-grained model of the Four C's, developed by Kaufman and Beghetto (2009). This model splits creativity hierarchically into Big-C, Pro-C, little-c, and mini-c creativities. Big-C creativity is the H-creative work of master creators who are eminent in their field. Pro-C creativity is the work of creative professionals which is not historically significant, yet is successful enough to provide for a creative career. Little-c creativity is the work of ordinary people, inventively solving problems in daily life or producing creative artifacts as a hobby. Mini-c, finally, refers to the creative work of children. Each of these forms of creativity may be evaluated in a different way. Note that some of the four C's imply something at work in the Press perspective: Big-C creative people are identified by their fame and influence, and Pro-C creativity depends on enough people buying a creative product to provide the creator with income. Without a Press to influence and reward them, all creative adults would be little-c creative.

Bartel (1985) objects to defining novelty as the transformation of conceptual space. He points out that one can imagine conceptual spaces being changed in trivial, random, or uninteresting ways. Bartel discusses the distinction between the terms "unique," "different," and "original," and proposes a definition of original works as those which are an *origin*. That is, a work is original if it is the first to display some unique or different attribute which is then adopted by other works. (Again, this implies that Product-based novelty is dependent on actions taken by the Press.) Merely novel works are not copied because the first such work already exhausts its own interesting possibilities. Note that, while this definition suggests H-creativity, it is not difficult to adapt to P-creativity; artists can and do copy themselves. Dasgupta (2011) defines A-creativity (standing for *antedecent*, any H-creative product which has never been seen before) and C-creativity (standing for *consequence*, an H-creative product which influences others). While A-creativity can be assessed immediately, assessing C-creativity requires a Press-based historical perspective.

Takeaways:

- If using novelty and value, consider what kind of novelty the product should have.
- Be careful to distinguish between meaningful novelty and randomness.

5.1.2 Value. The idea of value comes with questions of its own. Who decides if a work is valuable or not, and on what grounds? An obvious answer is that the wider culture, the Press, will decide. But it is easy to think of creative people (e.g., van Gogh or Mendel) who were ignored in their day, and assigned great cultural value posthumously. Does this mean that they became creative posthumously? Some artists have multiple phases of greater or lesser popularity after death. It is difficult to imagine that changes in the dead artist's actual creativity are behind these shifts (Weisberg 2015). But Csikszentmihalyi (1996) argues that a person's creativity can and does change after their death—because Csikszentmihalyi views creativity as a Press phenomenon, residing not in the creative person but in their interactions with the wider culture, which naturally changes over time, even when the creative person is dead. Dorin and Korb (2012) also bring up the example of cultural biases. Women's art has historically been overlooked due to systemic sexism. Does this mean that women's art is actually less valuable? Less creative?

Philosophers such as Gaut (2010) bring up the idea of "negative value." For example, imagine a terrorist who comes up with a new, unusual, and effective way of carrying out attacks. Is the terrorist creative? If so, who is assigning value to the terrorist's work? One solution is to deny that destructive acts are ever creative; another is to define them as valuable if they effectively serve the creator's goals. Cropley et al. (2008), who study this type of malevolent creativity, define acts as "subjectively benevolent" if they are beneficial to one person or group at the expense of another. In other words, a creative terrorist attack is assigned value by the group of other terrorists sharing the same political aims, even if they are not valuable to society at large. "Value" is sometimes phrased as "usefulness" or "appropriateness to the task," which accords with the intuition that even negative or pointless tasks could be done creatively (Kaufman and Baer 2012).

Weisberg (2015) argues that value should not be a criterion for creativity. In addition to artists becoming more creative after their deaths, a changeable definition of "value" means that it is difficult to track the validity of creativity research. One generation might consider Group A's work valuable, and Group B's work not valuable. A second generation might reverse these judgments, considering only Group B's work valuable. The previous generation's creativity research therefore becomes invalid, unless it can be shown that the conclusions drawn about Group A also apply to Group B. Weisberg suggests defining creativity as "intentional novelty."

Bown (2012) also argues against the use of value, citing processes like plate tectonics, which produce novel artifacts without a goal. Bown calls these processes "generative creativity." Humans arguably engage in generative creativity at times, such as when brainstorming. For some applications, such as a system that engages in brainstorming, judging the system by novelty alone may be appropriate. However, our view is that most applications do require value. A work of art or science is judged by its value when created by humans—so a computer system working in these domains should be held to the same standard.

Takeaway: If using novelty and value, define the specific audience for whom your system's products should be valuable.

5.1.3 Typicality and Ritchie's Model. Ritchie (2007) argues that typicality should be used rather than novelty. His argument is that, while novel output of some type by a computer is conceivable, the computer would first need to generate products that are actually of the given type. A novel string of incomprehensible letters would be easy to produce, but if that was all the computer produced, it would be difficult to argue that it is successfully writing poetry. Therefore, creative systems should reliably generate both typical and valuable output. In Ritchie's view, the most useful way to measure this would be to look at statistical properties of a system's output over time, and to compare them to the properties of an inspiring set—a set of exemplars the system uses to define its task. Ritchie lists many criteria that one could test for based on these statistical properties. For example, one might wish for the average typicality to be within a certain range, or for a certain proportion of the output to be above a given threshold for value. Which of these criteria, and what ranges and threshold values, are appropriate for a given system is an unsolved question. While typicality might appear to be the opposite of novelty, Ritchie's framing of them as separate criteria is supported by experiments such as Hekkert et al.'s (2003), who find that novelty and typicality in design both separately contribute to human preference.

Ritchie's model has fallen out of favor in recent years (Jordanous 2012b), but it has been used to evaluate real systems. Jordanous (2012b) found that Ritchie's model was cumbersome to implement, because of the need to give ratings to a large number of outputs. Its results are abstract and need to be rephrased to be useful to the researcher. Pereira et al. (2005) report that appropriate threshold values are usually unknown, and the criteria involving threshold values are very sensitive to changes in these values. Poorly chosen threshold values can invalidate other criteria by, for example, creating a scenario in which there are no valuable but atypical items. Defining typicality is challenging in practice, as it can be based either on content or structure; it is unclear what to do with, for example, an unoriginal artifact with structural errors.

Ritchie (2007) is ambivalent about attempts to use his criteria, stating that many of them are implemented incorrectly, particularly in the inspiring set. Many studies either choose an inappropriate inspiring set, or invalidate many composite criteria by basing typicality scores on similarity to the inspiring set, so that the typicality of generated work and the typicality of the inspiring set cannot be compared. RASTER, a thought experiment meant to demonstrate the insufficiency of Ritchie's criteria, similarly renders the criteria involving an inspiring set inapplicable (Ventura 2008). Pereira et al. (2005) found that this invalidation occurs when there are no atypical items in the inspiring set. RASTER also defines typicality and quality identically, which invalidates even more criteria (Ventura 2008). When not partially invalidated in this way, there may still be some value to Ritchie's model, and to the idea of typicality as a prerequisite to meaningful creativity.

Takeaway: Before testing novelty, consider testing if the system's output is appropriate for its intended domain.

5.1.4 Implementing Novelty and Value. If novelty and value are used to evaluate a system, then a specific procedure for their evaluation is desirable. It is especially desirable if novelty and value can be calculated by the computer itself—for use in the evaluation portion of a generate-evaluate loop (see Section 4.2)—although a computer system should not be judged only by its own self-assessment.

Pease et al. (2001) describe a number of possible methods for calculating novelty and value. For novelty, this includes difference relative to an inspiring set, level of transformation of conceptual space, relative complexity, membership in a fuzzy set based on an archetype, Bayesian improbability, and novelty perceived subjectively by humans. For value, it includes the emotional response of humans, or the degree to which the product achieves the creator's goal.

França et al. (2016) define novelty as Bayesian surprise, and value as *synergy*—a metric based on expert judgments of the value of pairs of components. Bhattacharjya (2016) defines quality in terms of a preference model which mathematically combines different aspects of a subjective judgment.

Elgammal and Saleh (2015) measure novelty in art history by creating a directed graph of influences based on semantic and perceptual traits—recalling Bartel's definition of originality as the creation of an origin—and found good correspondence between their system's behavior and the opinions of art historians.

Novelty and value can also be based directly on human opinion. Some researchers in this vein (Llano et al. 2014; Riedl and Young 2006; Veale 2015) have found that human assessments of novelty or surprise are negatively correlated with most other desired criteria. Indeed, in Llano et al.'s study, random statements are judged more surprising than either human or computational outputs, yet do poorly on every other criterion. Such "mere novelty" suggests that Ritchie was correct to require typicality as a prerequisite for novelty, but they are inconsistently reproduced. Some researchers supplement novelty and value with surprise—Boden's (1990) word for meaningful, significant novelty.

Takeaway: If you are judging based on criteria, consider operationally defining them so that the computer system can evaluate its own products and products-in-progress.

5.2 Other Criteria

Other product-based criteria exist besides novelty and value. Some are domain-specific. Cropley and Cropley (2005) define a hierarchical model specifically for "functional creativity"—the kind used in engineering. Their criteria are Effectiveness (the product solves the problem it was intended to solve), Novelty, Elegance (the product is pleasing, or goes above and beyond mere correctness, such as by being more cost-effective than previous solutions), and Generalizability (the solution can easily be applied to additional problems). Cropley and Cropley's model requires previous criteria in their list to be met before the next becomes relevant. Without novelty, an engineering product is not creative; but if a product is not effective, its novelty does not matter to an engineer. The remaining criteria of elegance and generalizability define higher levels of functional creativity.

Whether or not there is a similar hierarchy for non-functional creativity, remains unclear. On divergent thinking tasks, novelty and value often negatively correlate. Diedrich et al. (2015) found that, on such tasks, human judges' ratings of creativity correlated with novelty but not usefulness. Of course, this result may be specific to divergent thinking tasks, in which generating novel ideas without much evaluation of their usefulness is the whole point.

Other sets of criteria have been proposed. A popular set is Imagination, Appreciation, and Skill, the Creative Tripod (Colton 2008). However, subtleties in the thinking behind the Tripod mean it fits best with the Press perspective, and will be discussed in Section 6.1. Lehman and Stanley (2012) evaluate products by their Impressiveness: products are impressive when they are easy to appreciate, but difficult to recreate.

An extremely common technique is for researchers to use ad hoc domain-specific product criteria. The rationale for domain-specific criteria will be discussed in Section 7, and some examples given in Section 8.1. Ad hoc criteria are not ideal because their lack of clear theoretical grounding and testing, but the intuition behind them is sound. A given system may have the goal of producing work with certain traits, such as an elegant theorem, a poem fitting a given form, or a retweetable Internet meme. Criteria testing if these goals have been met can easily be added to a more general test of creativity, or if using novelty and value, criteria like these might be used to test value.

Takeaway: Appropriate criteria may depend on the specific domain.

5.3 The Modified Turing Test

A simple product evaluation without criteria is the modified Turing test: human subjects are given human-created and computer-created products, and are challenged to figure out which is which. If they cannot do it, then the computer system is creative. It is important to note that this is a *modified* Turing test. The original Turing test, for general intelligence, involves judges who can question the system however they like, while the modified test involves only static, finished products. Thus, it could not be said that a system which passed such an evaluation was intelligent, only that its products were indistinguishable from human products. However, the modified Turing test is frequently used in computational creativity (Schwartz and Laird 2015; Elgammal et al. 2017; Pearce and Wiggins 2001).

Pease and Colton (2011) criticize the modified Turing test on several grounds. First, the test encourages pastiche and superficial imitation. Second, the lack of interaction calls into question the modified test's validity. Third, the test limits access to framing information, such as how and why the product was created, which Pease and Colton consider an important part of evaluation. Finally, according to Pease and Colton, no current system can pass a Turing test, so evaluations are needed which can identify the strengths and weaknesses of systems even if they are only partially creative.

While we have some quibbles with Pease and Colton's arguments, we acknowledge that a modified Turing test is not a suitable evaluation method unless the system is specifically designed to create work similar to that of humans.

5.4 Consensual Assessment

Another criteria-free product evaluation has a much stronger pedigree. This is the Consensual Assment Technique (CAT) developed by Amabile (1983). The idea in consensual assessment is that accurate judgment of creativity is not formulaic but relies on the expertise of people who are recognized as creative in a domain. In a consensual assessment, human subjects create something—for example, a collage. The subjects should be matched for their level of experience with the chosen medium. A team of experts in the relevant field then looks at the products and each decides how creative each product is. An important point is that the experts are *not* given criteria such as novelty and value, or other instructions in how to make their judgments, except that they must use the full scale. They also must make the judgments on their own, rather than consulting with each other. If most of the experts agree (i.e., if they have good interrater reliability), the consensual assessment is successful. Even though the judges may be making different kinds of assessment, the CAT has good interrater reliability in practice as long as the expertise of the judges is high enough (Kaufman and Baer 2012). The use of experts regardless of their preferred method of judgment accords with how creative artifacts are evaluated in the real world, by critics, editors, peer reviewers, gallery owners, prize committees, and so forth (Kaufman and Baer 2012)—although, of course, jurors, critics, and editors do talk to each other in the process of judging. Like the modified Turing test, a computer cannot use the CAT to evaluate its own work, since it requires a group of expert assessors.

5.5 Product Evaluation, mk. II: Computational Aesthetics

In the arts, especially the visual arts, many researchers have attempted to create domain-specific criteria with a deeper perceptual grounding. This leads us to the field of computational and psychological aesthetics.

Galanter (2012) summarizes computational aesthetic research in visual art. Many of the aesthetic rules taught to humans in art schools, such as color theory, are difficult to encode in a way that matches human perception. However, many perceptual rules exist which are mathematical in nature, and can therefore be calculated. These include complexity, Zipf's law, JPEG compressibility, and prototypicality. Many of these do correlate at least somewhat with human preferences. Alternatively, a neural net can be trained to evaluate art without explicit rules. This need not be a binary classification. Systems such as DARCI learn to recognize many subjective qualities in artwork, such as "happy," "fiery," or "lonely," based on humans' descriptions (Norton et al. 2010).

How humans process visual art is a challenging field of study for human psychologists. Human responses to art can be highly individual (Dubnov et al. 2016; Juslin et al. 2016) and affected by context (Leder and Nadal 2014). Art processing involves the entire brain, rather than a specific art processing region (Leder and Nadal 2014). Still, psychologists are at work on theories outlining how this processing occurs.

One of the theories with the most explanatory power is the theory of representational fit (Sammartino and Palmer 2012). In this theory, images are preferred when they transparently reflect the work's intended meaning. In simple cases, this means humans will prefer images that are easy to process; but representational fit also explains complex and difficult artworks, such as ambiguous photographs. If ambiguity is the best way to make a point, and the audience understands this point, then they will appreciate the ambiguous artwork.

Alternately, Silvia (2005) suggests an emotional appraisal theory of aesthetics. If a human appraises an art object as novel and not understood, but believes themselves capable of understanding it, they will take interest in the art. Appraisal theory explains why novices prefer art that is easier to process, but also why art experts prefer more difficult work.

Leder et al. (2004) combine these and other theories into a model of the stages of aesthetic judgment. In Leder et al.'s model, the ultimate goal of looking at art, is "cognitive mastery," in which the audience feels they have figured out what the art means and what they think of it; different people will seek to achieve this mastery in different ways.

Art appraisal is also influenced by context. An artwork's price, title, setting, the artist's name, reported approval or disapproval of the artwork by other social groups, or even the viewer's current physiological state can influence human processing of art (Lauring et al. 2016). Therefore, a full accounting of human processing of visual art would also have to include contextual and bodily effects as well as social factors.

Similar work is available for other domains, often indicating that different domains have different aesthetic rules. Augustin et al. (2012), for example, find that humans value different qualities in visual art, film, and music. Jacobs (2015) identifies two distinct aesthetic processes that occur in literary writing—a "background" process which creates suspense or empathy through the situations that are described, and a "foreground" process which creates aesthetic appeal through the style in which they are described.

None of these theories can yet be encoded precisely enough for a computer. However, a researcher without intensive domain training might do well to refer to these or other theories of aesthetics when setting qualitative goals for the characteristics of their system's output.

6 PRESS PERSPECTIVE

After a person uses a process to create a product, the Press—other people—then receive it. The Press perspective studies how this reception occurs, and what kind of social effect a product needs to have to be called creative. To some extent, a Press perspective is necessary to every other perspective: the act of judging people, processes, or products is always done by people in a cultural context. (Or, if it is done by a computer, that computer was programmed in a cultural context by people.) The pure Press perspective is especially useful for researchers who want their system to make an impact on society, influence human opinion, and be recognized for its work.

In Rhodes' (1961) formulation, Press refers not only to "the press," as in cultural agents responding to a creative work, but to the general environment "pressing in" on the creative person. This includes social responses to creative products and the social causes of their creation. The social environment influences the values and beliefs of a creative person, determines what forms of education and training are available, and can direct the form of creativity through incentives and commissions (Csikszentmihalyi 1996). However, research from the press perspective has increasingly focused on responses and judgments.

Many scholars argue that creativity cannot be studied independently of its context. There are weak and strong versions of this claim. The weak version, advanced by Boden (1990), is that certain criteria are contextual. For example, H-creativity is contextual since it depends on what other people did and did not do before, and value is subjective because it depends on what people find valuable. Nevertheless, we could objectively study value by, for example, making a computer model of the values of a certain group and testing how well new works fit the model.

The strong version, advocated by Csikszentmihalyi (1999)—also known as a systems perspective—situates creativity entirely outside of person, process, and product. Since all judgments are subjective, in the strong Press view, it is epistemically and perhaps ontologically impossible to separate creativity from people's judgments of creativity. Therefore, creativity in practice is not separate from the people judging it. It is situated, not in the creator, but in an interaction between creator and audience. Even aspects such as skill and the ability to self-evaluate are in some sense internalizations of domain knowledge and field attitudes, which did not originate with that person (Csikszentmihalyi 1996).

Csikszentmihalyi breaks down some factors involved in Press-based creativity. These include the individual, domain, and field. The domain is the cultural and symbolic aspect of creativity: for example, what works have been produced in this genre, and what are its conventions and rules. The field is the social aspect: fellow creators, editors, curators, and critics who serve as gatekeepers. The word "genre" is used here not only to denote artistic genres, but also branches of science, business, activism, or any other creative activity. Even abstract tests such as the Torrance Tests have a domain (questions about chairs, etc.) and field (the scientists designing and scoring the tests). To be successful, in Csikszentmihalyi's view, a creative individual must change the way individuals in the field think, feel, or act (Csikszentmihalyi 1999). This means that Csikszentmihalyi is primarily interested in big-C, H-creativity, not in creativity's everyday forms (Csikszentmihalyi 1996).

The question of who performs the evaluation—of who constitutes the field—leads into the cross-cultural study of creativity. A review of cross-cultural studies by Lubart (1999) finds that not all human cultures agree on what constitutes creativity. He states that Westerners focus on originality and the production of something new, while Eastern cultures focus on the expression of insight, growth, and inner truth. He also describes differences in where creativity is thought to be situated. In Bali, for example, musical groups can be distinct from one another, but individual musicians within the groups may not. Therefore creativity occurs here only on the group level.

For computational creativity, Press evaluation is a challenge since most people are not used to seeing computers as creative. Some researchers study how exactly these people respond to creative artifacts made by a computer, or on effective methods for convincing them that the computer's work is worth consideration. These issues of perception will be discussed more fully in Section 8.2.

6.1 The Creative Tripod

Some researchers design systems with an explicit persuasive element, with the goal of convincing humans that the system is creative. Colton (2008) leads this trend. His Creative Tripod appears to be Person-focused at first: it presents several traits that a creative system should have, namely, Skill, Imagination, and Appreciation. Colton's assertion is not that a creative system must possess these qualities, but that a creative system must *appear to possess* these qualities. Much of Colton's research involves making machines more persuasive in convincing an audience that they have these creative qualities. This includes work on the framing of artifacts, taking advantage of contextual effects that can sway human judgment (Charnley et al. 2012).

Certain objections to the Creative Tripod have been raised. Colton et al. found that, to overcome skeptics' objections to computational creativity, the original three criteria were not enough. They added the new criteria of Learning, Intentionality, Accountability, Innovation, Subjectivity, and Reflection (Colton et al. 2014). Not much justification is given for these specific additions, except a statement that they addressed the most common objections to characterizing a system as creative.

Bown (2014) objects to the Tripod because the terms are not given clear definitions; therefore, they "cannot be distinguished from trivial pseudo-versions of themselves." A system can be argued to possess imagination, for example, if the programmer put something into it which the programmer believes is related to imagination—which bypasses any falsifiable inquiry into whether this actually counts as imagination. Many papers perform exactly this kind of trivial evaluation, and even obviously uncreative systems can appear to pass the Tripod with the right argument. For example, Ventura's thought experiment, RASTER, generates the pixels of images at random, and outputs the images if a similar image can be found online. Ventura (2008) describes RASTER as meeting Tripod criteria: *imagination* because it engages in random search, *appreciation* because it uses a (simplistic) fitness function, and *skill* merely because it produces images.

Smith et al. (2014) give their own working definitions to the Tripod. Skill is the ability to produce something useful; imagination is the ability to search the conceptual space and produce something novel; appreciation is the ability to self-assess and produce something of worth. (Note the implied links between the Tripod and a Product perspective, as well as Boden's work.)

Jordanous (2016b) identifies Colton's tripod with three aspects of the SPECS model: Skill to *domain competence*, Imagination to *variety, divergence, and experimentation*, and Appreciation to *thinking and evaluation*. Jordanous (2012b) argues that evaluating work based on the tripod requires process information about the system's behavior over time. Product, Person, and Process concepts frequently creep into Press evaluation, because the humans reacting to a creative system are assumed to be using these concepts themselves.

There is nothing wrong with the Creative Tripod criteria when they are defined in this way. However, inappropriate uses of the Tripod serve as a cautionary tale: for any criteria, Press or otherwise, if we have not specified just what we mean by each of our criteria, our evaluation becomes meaningless.

Takeaway: Make sure to define any term used as a criterion for evaluation before the evaluation begins.

6.2 Measures of Audience Impact

A major Press idea is that a creative product should have an impact on its audience. Several researchers are interested in measuring audience impact. These include Colton et al. (2011), who describe the interaction between a creator's work and an idealized audience with the IDEA model: Iterative Development, Execution, Appreciation. Colton et al. describe development moving through stages based on how novel it is, from completely derivative work to humanlike work to work so novel that humans cannot comprehend it. Audience impact at any stage is measured with two variables: *wellbeing* (how much the audience likes the work) and *cognitive effort* (how prepared the audience is to spend time trying to understand it).

Like Ritchie's criteria, these variables can be combined in a variety of ways. For example, a work with a high standard deviation in wellbeing would be considered "divisive." IDEA is a descriptive model, and it is up to the researcher to decide which of the possible adjectives applied to their system are desirable.

Burns' (2015), EVE' model measures the mental processes of the audience in another way, defining creativity as surprise with meaning. In the EVE' model, which has been applied to jokes, simple visual art, advertisements, and poetry, an expectation (E) is set up, then violated (V). The violation is accompanied by a new explanation (E') which accounts for the unexpected events. This explanation may be overt, or may happen implicitly as the audience retrieves contextual information from long-term memory (Dubnov et al. 2016). If the audience is surprised by a work, *and* can make meaningful sense of it, they approve. In experiments, ratings of surprise multiplied by ratings of meaning accounted for 70% of the variability in ratings of creativity (Burns 2015).

There is some theoretical support for the EVE' model. As Dubnov et al. (2016) point out, it is compatible with the appraisal theory discussed in Section 5.5. A surprising stimulus is appraised by an audience as novel and not understood, and the subsequent explanation causes the audience to appraise themselves as able to understand it. According to Jacobs (2015), surprise followed by explanation is what constitutes foregrounding in literature. In a longer work, oscillation between surprise and explanation is constant. However, it remains to be seen experimentally if the EVE' model can be applied straightforwardly to creative works with longer, more complex content.

Depending on the domain, any of these methods may be appropriate for testing if a creative system is having the desired effect on its audience.

Takeaway: Decide what effect the system is intended to have on its audience.

6.3 Interactive Art

Bown (2014) suggests that computational creativity researchers should perform evaluation through the lens of interaction design. The designers of a creative system must consider how the audience will interact with their system and what effect they wish it to have. Fortunately, there is already a great deal of research on creative interaction design in the domain of interactive art, studied by human-computer interaction researchers, museologists, and others.

Candy and Bilda (2009) describe three types of audience engagement: immediate (catching attention), sustained (attending to the art for a period of time), and creative (having a lasting effect that somehow changes the audience). Similarly, Bollo and Dal Pollozo's (2005) model of museum exhibits involves variables such as Attraction (the percentage of visitors who look at an exhibit) and Holding Power (the amount of time an average visitor spends looking at the exhibit). Edmonds et al. (2006) relate these traits to parts of a display. Attractors increase attraction, Sustainers increase holding power, and Relators encourage the visitor to keep thinking about the artwork and return later.

HCI researchers also make use of qualitative, descriptive methods in assessing audience response. Her et al. (2014) review many of these methods.

All of this work is helpful for researchers crafting an interactive system. It is more difficult to apply theories from HCI to a system that creates something static, such as a painting or poem. It is even more difficult to apply them to a system that performs scientific, mathematical, or some other form of creativity in which the goal is to soberly present a theory that experts recognize as meaningful. However, if interaction is one of a researcher's goals, then care should be taken to consult the literature on interaction which already exists.

Takeaway: If your system is interactive, evaluate it according to the standards of interaction design.

6.4 Creativity Support Tools

Another application of interaction design is to co-creativity or creativity support tools. Rather than producing an artifact by themselves, these tools make it easier for a human to be creative. It is possible to evaluate a creativity support tool using any of the four perspectives, but the method that usually makes sense is to evaluate the quality of the user's interaction with the tool.

Evaluation of creativity support tools is well studied, and includes empirically validated rating scales measuring the extent to which a tool supports a particular creative goal (Carroll and Latulipe 2009). Apart from these scales, evaluation of co-creativity tools can focus on usability, enjoyability (Kantosalo et al. 2015; Waller et al. 2009), usefulness to creative professionals (DiPaola et al. 2013; Kantosalo et al. 2015), or the quality of output created using the system (Lee et al. 2016; Shibata and Hori 2002; Waller et al. 2009). Kantosalo and Toivonen (2016) also classify co-creativity systems as *alternating* (where the human and computer take turns modifying an artifact) or *task-divided* (where the human and computer are responsible for different subtasks).

Takeaway: If your system is co-creative, evaluate it using existing standards for creativity support tools.

6.5 Artificial Social Systems

In Section 4.2, we mentioned Glăveanu's (2015) comment that the evaluation phase of the human creative process is based in perspective-taking. Now that we have seen Csikszentmihalyi's theory, we can be clearer about the humans whose perspectives are important: they constitute the field. Several researchers have asked what happens if creative systems serve as each other's field. They have created multi-agent systems, either of robots or of software modules, which influence and learn from each other's work (Kirke and Miranda 2013; Linkola et al. 2016; Saunders et al. 2010). Corneli et al. (2015) imagine how a computer could go through a writing workshop, in which drafts are critiqued by other computers trained on similar tasks. Systems also exist which simulate groups of musical performers improvising together (Eigenfeldt et al. 2017; Puerto and Thue 2017).

6.6 Cultural Success

A final method for press evaluation is to publish the creative product as a human would publish theirs. In art, this means submitting the work to an art exhibition or gallery. It is not uncommon for HCI researchers creating interactive art to do exactly this. The artwork is judged a success due to opinion of the audience, gallery admissions, statements of interest by curators, or whether the artist was invited to submit work to further exhibitions (DiPaola et al. 2013; Sheridan et al. 2005; Tresset and Deussen 2014).

An artwork's success with these gatekeepers can be a useful definition of press-based creativity; it naturalistically reflects the metrics most working human artists use to judge their own success. Nor is this a method restricted to art. Scientific advancements are evaluated through peer review and publication; entertainment for a general audience is evaluated by its commercial success; Internet memes are evaluated by how often they are shared.

Jordanous has quantified some of these forms of success, calculating a musician's cultural value by the number of comments they receive on a digital music site (Jordanous et al. 2015) and a computational creativity researcher's impact by the number of non-self citations in the 5 years following publication (Jordanous 2016b). Interestingly, the number of non-self citations of a system "roughly aligns" with expert judgments of the system on metrics like Ritchie's model or the Creative Tripod—but not with the judgments of experts who were asked "how creative is this?"

Cultural success of is the ultimate form of Press evaluation, and is very similar in principle, if not in methodology, to the Consensual Assessment Technique (Section 5.4). Cultural success is a goal that takes a creative product seriously in its entirety—and that subjects it, however indirectly, to the same career pressures that would be applied to a creative human.

Takeaway: If your system is meant to create artifacts similar to those of humans, consider submitting it to the same cultural gatekeepers who would judge a human artifact and studying their response.

7 ARGUMENTS AGAINST EVALUATING CREATIVITY

Now that we have looked at all four perspectives, it is time to mention some important counterpoints from researchers who question whether creativity evaluation is possible at all.

7.1 Domain Specificity

One surprising argument against the existence of creativity is Baer's (2012) theory of domain specificity. Baer states that there is no such thing as creativity—or, rather, that there are many creative skills, but there is no underlying process which informs them all. Being creative in one domain does not imply the ability to be creative in other domains; therefore, to call a person or process creative without specifying the domain is not scientific. Baer writes,

> "It is sometimes useful to group together beautiful artifacts, fascinating ideas, brilliant designs, and ingenious theories and call them all creative, but that does not mean that they share any underlying unity."

Baer's theory, while disheartening, is supported by plentiful evidence. The Consensual Assessment Technique requires domain-specific experts, and a subject's CAT results in one domain do not significantly intercorrelate with their results in another domain. The only times there have been even modest correlations are in highly related domains, like different kinds of stories or visual art (Baer 2012).

Similarly, the Torrance Tests predict creative achievement only in their associated domains; the verbal Torrance Test does not correlate with the figural Torrance Test (Baer 2012). Such tests correlate only modestly, and sometimes questionably, with actual creative achievement either at the time of taking the test or later in life (Baer 2011). Personality-based correlates of creativity are also different across domains (Baer 2012). It is difficult to explain such a lack of correlation without acknowledging that some component of creative thinking is different for each domain.

Some results from previous sections support domain specificity. Readers will recall Augustin et al.'s (2012) discovery that different product-based qualities are important in different artistic forms. Mace and Ward (2002) and Bourgeois-Bougrine's (2014) studies of working artists, while roughly parallel, showed differences in process, especially in how a work is completed. Product-based theories are frequently domain specific, such as Cropley and Cropley's (2005) theory of functional creativity, or Jacobs' (2015) theory of foregrounding and backgrounding in literary work. Press-based HCI studies of interactive gallery artwork are often difficult to apply to any other domain.

There is also some evidence that aspects of creativity are domain general. Eminent human creativity often involves cross-applying knowledge from one domain to another (Csikszentmihalyi 1996). Jordanous (2012b) found that different SPECS criteria are considered important in different domains, but 4 of the 14 criteria are very important for every domain: Generation of Results, Originality, Spontaneity and Subconscious Processing, and Value. Since "novelty" and "originality" are synonyms, this would imply that the "novelty and value" definition of creativity, and perhaps others, are domain general. Jordanous (2012b) thinks of creativity as partly domain general and partly domain specific.

The evidence for at least partial domain specificity is very strong. Computational creativity researchers must take note of this, and ensure that evaluation techniques are appropriate to the domain. If possible, generalized evaluation techniques should be tested for their applicability to specific domains. However, it does not follow that there is no creativity evaluation. Researchers can continue to evaluate systems that create art, music, mathematics, and so forth—on the understanding that the evaluations for these different systems will also be different.

Takeaway: Make sure that any evaluation technique used is appropriate to the creative domain.

7.2 Other Arguments

Beyond domain specificity, there are other arguments against measuring general creativity. One such argument is that creativity should not be quantified. Boden herself (1990) takes this angle, preferring to ask "what parts are creative and why?" rather than "how creative?", as this produces more nuanced information for the system's creators. Nake (2012) argues that the quantification of creativity is an American invention, and risks commodifying creativity by framing it as an object one must have a certain amount of, rather than a quality that emerges in a social context. However, many computational creativity evaluations are not quantitative; SPECS is purely qualitative, for instance (Jordanous 2012b).

Related to this is the argument that *computational* creativity should not be measured by *human* standards. Loughran and O'Neill (2016) typify this argument, stating that humans can already produce creative artifacts which are pleasing to humans. To Loughran and O'Neill, it is more interesting to see what computers produce according to their own, non-human standards. However, most advances in computational creativity will require at least some evaluation: otherwise it is difficult to show that an advance has been made. If one wants to judge computers by non-human standards, one can still make statements about the computer's success performing to these standards, perhaps by using an autonomy-based model such as Guckelsberger's (2017) or a categorization like Ventura's (2016) of the tasks the computer takes on.

Finally, some say creativity is inherently human and can never be present in computers. Boden (1990) lists lines of argument here, as does Minsky (1982).

First, there is the argument that human creativity is an inexplicable gift which cannot be modeled computationally. Minsky (1982) convincingly refutes this argument, stating that creativity seems to be a combination of ordinary cognitive processes. Collecting domain knowledge, generating ideas, evaluating them, and revising can all be in principle done by a computer, as can the progression from an idea to a plan to a finished implementation. The luck and social factors that lead to a creative achievement being accepted by its field (Csikszentmihalyi 1996) are also not impossible for a computer, assuming that the field is open to the possibility of a computer achieving something.

If we accept for the sake of argument that human creativity is somehow not computational, we can still get useful results from computational creativity by focusing on product or press. If computers produce believable creative work and influence human culture, then they are doing something useful regardless of their means of doing so. Furthermore, computer systems can still provide evidence for and against hypotheses about parts of the process of human creativity (Boden 1990).

Second, there is the argument that, due to a computer's lack of richly embodied life experience, its performance will never match that of the greatest creative humans. Boden (1990) agrees with this, but says there are many other reasons why modeling creativity with computers is useful. Alternatively, several other have proposed solutions to this problem. Colton (2008) lists it as a reason for providing framing information, to create the illusion of humanlike experience. Moreover, a number of creative systems are actually embodied, either as robots (Infantino et al. 2016; Tresset and Deussen 2014) or in a virtual world (Aguilar and Pérez y Pérez 2014).

Third, there are arguments that even if a computer had humanlike processes and products, it could not be "really" creative (Boden 1990). These arguments can stem from appeals to the non-biological nature of computers, or to variants on Searle's (1980) Chinese Room argument, or to a lack of consciousness on the computer's part, or finally, to the simple belief that creativity is a property only of humans. Chinese Room-style arguments can be countered by the argument that understanding is an emergent property of the system as a whole, or Minsky's (1982) argument that humans do not "really" understand things either: our commonsense knowledge consists of imprecisely defined concepts induced from sensory perception, and it is possible for a computer to do this sensory processing as well. As for consciousness, Minsky (1982) argues that this is merely an ability to monitor oneself. Many creative systems do have a rudimentary ability to monitor their own work, and in principle, nothing stops this self-monitoring from becoming more sophisticated. Linkola et al. (2017) discuss one possible framework for this kind of self-monitoring.

Arguments about "real" creativity, "real" understanding, and "real" consciousness can be pernicious due to the ill-defined nature of these terms. This lack of definition leads to moving goalposts and unfalsifiable arguments. McCormack and d'Inverno (2014) suggest that, once a computer can perform a task, humans will no longer see that task as creative—even when humans do it. As a result, creative humans will concentrate on whatever tasks computers have not yet achieved. There is some precedence for this in, for example, the movement away from photorealism in painting following the invention of the camera.

In our opinion, none of these arguments are reasons to do away with evaluation. However, researchers should be aware that these viewpoints exist. For some members of the audience, "real" creativity may be a moving goal that a computer, no matter how sophisticated, can never quite meet.

8 ISSUES IN COMPUTATIONAL CREATIVITY EVALUATION

So far, we have seen many theories of creativity and methods for evaluating creativity. We have seen potential issues arise, such as debates about definitions of terms, a lack of autonomy in existing systems, the cultural specificity of many judgments, and the potentially domain-specific nature of creativity. Some of these issues have obvious implications when applied to evaluation in practice. Others are unsolved problems.

We now turn to some issues that arise due to the practicalities of computational creativity evaluation. Some of these have to do with problematic evaluation methods, or ones not well supported by theory and evidence. Some are practical questions to think about, while others are common pitfalls. Discussing problematic evaluations necessitates discussion of meta-evaluation: how to evaluate evaluation methodologies.

8.1 Implementations of Models and ad hoc Tests

It is one thing to propose criteria for creativity, and another to operationalize them to be used in practice. A number of researchers have operationalized models such as Ritchie's criteria (Gervás

2002; Tearse et al. 2011) or the Creative Tripod (Chan and Ventura 2008; Monteith et al. 2010; Norton et al. 2010; Smith et al. 2014), either by creating a questionnaire based on the model's criteria, or by somehow automating the judgments. Others have created their own questionnaires ad hoc. Operationalizations of novelty and value (Section 5.1.4), Ritchie's criteria (Section 5.1.3), and the Creative Tripod (Section 6.1) have already been discussed; we now turn to the issue of questionnaires which are not, or only partly, based on such models.

Some researchers evaluate only part of a model, or combine criteria from multiple models. Karampiperis et al. (2014) combine novelty and surprise (but not value) with Lehman and Stanley's (2012) impressiveness. A few systems are designed specifically for the Appreciation portion of the Creative Tripod (Norton et al. 2013).

Questionnaires used to evaluate creative systems do not necessarily adhere to an established model. Neither do the criteria used in a system's internal fitness measures, if applicable. More commonly, researchers evaluate systems according to ad hoc criteria. We have encountered dozens of such criteria, and Jordanous' (2012b) meta-analysis describes them as one of the most common forms of evaluation, but for space reasons, we make no attempt to list them all. A few illustrative examples of domain-general ad hoc criteria are meaningfulness (Das and Gambäck 2014), interestingness (Román and Pérez y Pérez 2014), and coherence (Harmon 2015); while domain-specific ad hoc criteria include grammaticality for poetry (McGregor et al. 2016); whether a respondent would use a generated image as desktop wallpaper (Norton et al. 2013); and induction of physiological responses by music (Monteith et al. 2010). Many of these studies include both domain-general and domain-specific criteria, as well as novelty, value, surprise, or a modified Turing test.

Some of these criteria are based on a sophisticated theory, as in the example of physiological responses. However, more often, there is little or no justification for why these criteria were used, except that they were the criteria the researcher happened to be interested in. Arguably, since part of creativity is likely to be domain specific, ad hoc criteria which are domain specific may be a better fit for a given project than standardized criteria. However, if a researcher argues that their system is creative—as opposed to merely retweetable or humorous—then such criteria must be based on experimental evidence with regard to the system's domain and rigorous theory.

Both ad hoc and theoretical criteria should be tested for properties such as construct validity, in the same way as other psychological criteria. The only model we are aware of that has been validity tested in this way is Carroll and Latulipe's (2009) Creativity Support Index, which is used for creativity support tools and is inappropriate for non-co-creative systems.

Some preliminary results indicate that most creativity criteria are not independent. Tapscott et al. (2016) found that measures of quality and narrative potential are interdependent; Pereira et al. (2005) found that ratings of typicality and quality under Ritchie's model are not independent. Lamb et al. (2015) found similar results with Ritchie's model, the IDEA model, and the Creative Tripod when assessed by nonexperts. A model in which criteria are demonstrably orthogonal to each other has yet to emerge, meaning that we do not have a sense of any "dimensionality" that might underlie creativity.

A related problem is that some criteria are operationalized simplistically; defining "poeticness," for example, as the use of a meter and rhyme scheme (Das and Gambäck 2014), when the majority of contemporary English poetry does not rhyme. Beginning with such goals in the early stages of a project is defensible, but for a system's output to be taken seriously, effort must be made to evolve toward criteria that resemble the actual criteria applied to human work.

Takeaway: Whenever possible, if you are using criteria, use criteria for which there is existing experimental or theoretical evidence.

8.2 Opinion Surveys, Non-expert Judges, and Bias

Aside from the criteria on a questionnaire, several other issues can arise. One issue is rater expertise. Intuitively, people who know little about a kind of creative artifact might not be good judges of those artifacts. Some researchers have expressed this intuition in their published work: for example, Gervás and Veale both independently worry that humans will rate their systems' output too highly because they do not understand it (Gervás 2002; Veale 2015). The intuition is supported by considerable evidence. The Consensual Assessment Technique requires domain expert judges for a reason: only domain experts can be trusted to judge artifacts in that domain (Kaufman and Baer 2012). Non-expert judges lack interrater reliability. Even when they agree, the validity of their judgments is in serious question, because they fail to correlate well with the judgment of experts.

This should not be surprising; it is well known in cognitive science that experts perceive the subject of their expertise differently from novices, "chunking" and analyzing patterns that are invisible to a novice (Gobet and Simon 1998). In art evaluation, experts evaluate art differently from novices in several ways. Photography experts carry out less simplistic evaluation than non-experts, and prefer more unfamiliarity and uncertainty (Galanter 2012). The relationship between novelty and typicality is different for experts and non-experts, with experts showing a stronger preference for novelty (Hekkert et al. 2003). Art experts focus more than novices on relationships between properties, and less on the properties themselves; they rely more on domain-specific knowledge, while novices rely more on their general life experience (Kim et al. 2011). Art experts show less pronounced emotional responses to visual stimuli than novices, rely less on emotion in their judging of artworks, and are more tolerant of negative emotions in art (Leder et al. 2014). Experts and novices even move differently, with novices either losing interest or hovering around the art without a mental framework for interpretation (Ryokai et al. 2015). Rather than being a weaker version of expert judgment, novice judgment tends not to correlate with expert judgment at all (Kaufman et al. 2008) and can even run in the opposite direction (Lamb et al. 2015).

Non-expert raters in computational creativity face an additional problem: they typically do not know how to apply the concept of creativity to a machine. As mentioned earlier, many people are reluctant to attribute creativity to machines; and experiments suggest that those who are willing are frequently unsure how to do it. Jordanous (2014) found that participants in a survey asked how creative a system was, expressed confusion as to what definition of creativity to use, and admitted they were likely to conflate creativity with other factors. Norton et al. (2011) found that art students were reluctant to evaluate a computer's creativity without knowing more about its process.

Some researchers (Colton 2008) worry that this reluctance will lead to bias against creative computers. The evidence for such bias is patchy. Moffat and Kelly (2006) found that musicians and non-musicians are biased against computer-generated music, but their sample size is quite small, and other ways of analyzing their data did not yield this result. Other researchers have generally not reproduced Moffat and Kelly's results. Friedman and Taylor (2014), Norton et al. (2015), and Pasquier et al. (2016) found that, while individuals differ, there is little overall bias against computational creativity in the general population. However, the idea of bias against computers is still widely cited. Colton et al. (2008) recommend the use of framing information to combat bias, but McGregor et al. (2016) found that framing information does not significantly affect human ratings of computer-generated artifacts. A more subtle question is if humans are biased toward familiar and humanlike forms of creativity. Guckelsberger et al. (2017) raise this question when discussing embodied creative agents whose bodies might be very non-anthropomorphic. Framing information might be useful to help humans understand an agent who is unlike them; but this remains to be tested.

Some researchers argue against opinion surveys altogether. Colton (2012) worries that evaluating systems by surveying groups of humans would lead to "creativity by committee"—which would presumably be bland or otherwise undesirable. However, no real evidence underlies this claim, and some evidence—the success of the CAT, for instance—suggests that groups of experts can evaluate competently.

Jordanous (2014) argues that a lack of reliability renders opinion surveys unsuitable as an evaluation method; she recommends a detailed evaluation by a single expert. However, Jordanous's argument is not as general as it seems: the "opinion surveys" she cites are based on asking the question "how creative is this?" to mixed expert and non-expert raters (Jordanous 2012b). It does not follow that the same question asked in a structured way to experts (as in the CAT), or broken down into components (as in SPECS), will not provide suitable results. However, in any case involving human judges, care must be taken to avoid any problems caused by confused, inattentive, biased, or inexperienced judges.

Takeaways:

- Use expert raters whenever possible, and do not statistically lump experts and non-experts together.
- Provide guidance to raters who are not used to judging a computer's creativity, but do not assume that they will be biased against the computer.

8.3 Meta-evaluation

Some researchers have turned to the question of meta-evaluation—that is, if there is a systematic way to judge the merits of evaluation techniques.

An early study in meta-evaluation was Pearce et al.'s (2002) for music generation. Meta-analyzing papers in that field, they discovered a "methodological malaise." Most researchers neither clearly specify a motivation nor choose an appropriate evaluation method.

According to Pearce et al. (2002), different purposes necessitate different kinds of evaluation. The purpose of a creative system might be to create art: that is, the developer wishes to express themselves using a computational system. Pearce et al. argue that, while there is nothing wrong with this, it is art and not science. It should not be published in a scientific journal unless the artist makes a scientific or technological advance in the course of their work. The art itself should be evaluated through Press: critical acclaim, popular appeal, placement in curated exhibits, and so forth. Indeed, as we noted in Section 6.6, these methods are used by many.

Another purpose might be the creation of a general-purpose creative system, either autonomous or with human collaboration. This is an engineering task, and should be subject to a normal engineering process, including requirements analysis, specification, and testing. Researchers often fail to specify their engineering goals—in particular, to list practical scenarios in which their system would be useful, and to state the conditions under which they will deem the system successful.

The two other purposes cited by Pearce et al. (2002) are scientific: investigating a theory about an artistic matter (for example, a theory of musical style) or investigating a theory about the cognitive processes of human artists. These should be investigated using the scientific method: clearly stating a hypothesis, using methods derived directly from theory while minimizing confounding factors (e.g., one should not go into the system and make ad hoc manual tweaks so that it sounds better), and systematically attempting to disprove the hypothesis. Subjective evaluations by the researcher, which are generally not falsifiable, should be avoided (Pearce et al. 2002).

Jordanous (2014) proposes a set of five standards for evaluation methodologies. Good evaluations should be correct, in the sense of accurately and comprehensively portraying a system's creativity. Jordanous does not believe in ground truth about creativity, but correct feedback should be appropriate and realistic for the system being evaluated. Good evaluations should be useful for understanding and improving the system. They should faithfully capture creativity (as opposed to some other trait which could be conflated with creativity). They should be usable and easily applied. Finally, good evaluations should generalize across various types of creative system (note that this contradicts Baer's insistence on domain specificity). Jordanous states that a single methodology that works for every system probably does not exist; instead, we can use the five standards to talk about the strengths and weaknesses of a methodology and its suitability for a particular purpose. An exercise for the reader might be to choose some evaluation techniques discussed in this article, and perform one's own informal analysis of their merits based on these five standards.

Takeaway: Be clear about how you conceptualize your system—as art in itself, as an engineering product, or as a scientific experiment—and choose an evaluation which is appropriate to the chosen paradigm.

9 CONCLUSION: IMPROVING THE ASSESSMENT OF CREATIVITY IN COMPUTATIONAL SYSTEMS

We have now seen the major theories from the four perspectives of what creativity is; the major ideas from the four perspectives about how to evaluate creativity; some counter-perspectives proposing that creativity is not one thing, or perhaps should not be evaluated at all; and some additional pitfalls that occur when evaluating computational creativity in practice. A reader of this article is now equipped to think about how to evaluate their (or another researcher's) creative computational system in an evidence-based way.

To summarize, our suggestions for best practice are as follows:

For *person:*

- Consider this perspective if you want your system to be viewed as a general creative agent because of its inherent traits.
- If your system is meant to exhibit specific human cognitive traits, use existing tests to measure those traits. Otherwise, claims about your system should likely be based on another perspective.

For *process:*

- Consider this perspective if you want to model human creativity, or you want to make an argument that your system is creative because of the kinds of tasks it does.
- Think about how your system explores its conceptual space. What are the parameters of the space? Do you want it to combine existing ideas, to generate something radically transformative, or to explore the space more modestly?
- Consider building your system based on a generation-evaluation loop or a more complex looping process. The evaluation stage in the loop should lead it to incrementally improve.
- Consider building your system to move from inspiration to planning to creation rather than trying to generate a full artifact all at once.
- Decide on a level of autonomous decision making that it is realistic for your system to have at its current stage. Ensure that you understand which decisions are made by the computer and which by humans, and that your description of the system realistically represents this. Consider increasing your system's autonomy in later versions.
- Process evaluations tend to be either a placement of the system in one of several broad categories, or a qualitative analysis of the system's process strengths and weaknesses.

For *product:*

- Consider this perspective if your system has the goal of producing something useful to humans.
- If you are judging based on criteria, consider operationally defining them so that the computer system can evaluate its own products and products-in-progress.
- If using novelty and value, consider what kind of novelty the product should have.
- Before testing novelty, consider testing if the system's output is appropriate for its intended domain.
- Be careful to distinguish between meaningful novelty and randomness.
- If using value, define the specific audience for whom your system's products should be valuable.
- Appropriate criteria may depend on the specific domain. In an artistic domain, if empirical studies of the aesthetics of your domain exist, consider basing your criteria on these.
- The modified Turing test is not appropriate unless your specific goal is to imitate existing human work. Product-based tests in which creativity is specifically judged by experts, such as the CAT, have more validity.

For *press:*

- Consider this perspective if you want your system to make an impact on society, influence human opinion, and be recognized for its work.
- Decide what effect the system is intended to have on its audience.
- If your system is interactive, evaluate it according to the standards of interaction design.
- If your system is co-creative, evaluate it using existing standards for creativity support tools.
- If your system is meant to create artifacts similar to those of humans, consider submitting it to the same gatekeepers who would judge a human artifact and studying their response.

For all perspectives:

- Avoid evaluating your system through rhetorical argument or other unfalsifiable techniques.
- Make sure to define any term used as a criterion for evaluation before the evaluation begins.
- Make sure that any evaluation technique used is appropriate to the creative domain.
- Whenever possible, if you are using criteria, use criteria for which there is existing experimental or theoretical evidence.
- Use expert raters whenever possible, and do not statistically lump experts and non-experts together.
- Provide guidance to raters who are not used to judging a computer's creativity, but do not assume that they will be biased against the computer.
- Be clear about how you conceptualize your system—as art in itself, as an engineering product, or as a scientific experiment—and choose an evaluation which is appropriate to the chosen paradigm.

The question arises as to why it is so common for researchers to deviate from best practices: for instance, to survey nonexpert judges, perform modified Turing tests, or make rhetorical arguments, when there are substantial reasons to distrust these methods. One common reason is time and effort. Many theoretically sound methodologies, such as the CAT, are too complicated to always be practical. Even the use of expert judges requires time, effort, and expense, as experts willing to perform such evaluations can be difficult to find. In an experimental field—particularly when making computational art that does not exactly resemble human art—it may be difficult even to identify an expert. If sheer difficulty causes the disconnect between theory and practice, then new ways of making theory simpler to implement are desperately needed.

Another interpretation is that the goals of practical researchers diverge from those of theorists. For example, if one's goal is not to construct a "really" creative system but to convince the public that one's system is creative, then rhetorical argumentation and the use of non-expert raters may be very appropriate. More seriously, it may be appropriate to survey non-experts about one's system if the system's goal is to entertain non-experts. Similarly, ad hoc surveys can be defensible on the grounds of domain specificity (Baer 2012). Apart from the CAT, existing evaluation methodologies are not domain specific. And as we have seen in Section 8.1, many ad hoc surveys used by computational creativity researchers contain domain-specific criteria. It may be that researchers avoid or modify general theories of creativity because they already understand that their task is domain specific.

If this is the case, then what is needed is a proliferation of more rigorous methodologies appropriate to specific domains. If a researcher needs criteria specific to the needs of music, or mathematics, or Internet memes, then their need must be matched by systematic attention to the problem of which criteria represent the specific needs of that domain. Such criteria can, one hopes, be based in existing scholarship pertinent to the domain in question, and solidified through experiments.

REFERENCES

Wendy Aguilar and R. Pérez y Pérez. 2014. Criteria for evaluating early creative behavior in computational agents. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, Ljubljana, Slovenia, 284–287.

Teresa Amabile. 2012. *Componential Theory of Creativity*. Harvard Business School, Boston, MA.

Teresa M. Amabile. 1983. The social psychology of creativity: A componential conceptualization. *Journal of Personality and Social Psychology* 45, 2 (1983), 357.

M. Dorothee Augustin, Claus-Christian Carbon, and Johan Wagemans. 2012. Artful terms: A study on aesthetic word usage for visual art versus film and music. *i-Perception* 3, 5 (2012), 319.

John Baer. 2011. How divergent thinking tests mislead us: Are the torrance tests still relevant in the 21st century? The division 10 debate. *Psychology of Aesthetics, Creativity, and the Arts* 5, 4 (2011), 309.

John Baer. 2012. Domain specificity and the limits of creativity theory. *The Journal of Creative Behavior* 46, 1 (2012), 16–29.

Christopher Bartel. 1985. Originality and value. *British Journal of Aesthetics* 25 (1985), 169–184.

Monroe C. Beardsley. 1965. On the creation of art. *Journal of Aesthetics and Art Criticism* 23, 3 (1965), 291–304.

Tarek R. Besold. 2016. The unnoticed creativity revolutions: Bringing problem-solving back into computational creativity. In *Proceedings of the AISB 3rd International Symposium on Computational Creativity*. CRC Press, Taylor & Francis Group, Sheffield, UK, 1–8.

Debarun Bhattacharjya. 2016. Preference models for creative artifacts and systems. In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, Paris, France, 52–59.

Margaret A. Boden. 1990. *The Creative Mind: Myths and Mechanisms*. Psychology Press, Hove, UK.

Alessandro Bollo and Luca Dal Pozzolo. 2005. Analysis of visitor behaviour inside the museum: An empirical study. In *Proceedings of the 8th International Conference on Arts and Cultural Management*, Vol. 2. International Association of Arts and Cultural Management, Montreal, Canada, Article 28, 13 pages.

Samira Bourgeois-Bougrine, Vlad Glaveanu, Marion Botella, Katell Guillou, Pierre Marc De Biasi, and Todd Lubart. 2014. The creativity maze: Exploring creativity in screenplay writing. *Psychology of Aesthetics, Creativity, and the Arts* 8, 4 (2014), 384.

Oliver Bown. 2012. Generative and adaptive creativity: A unified approach to creativity in nature, humans and machines. In *Computers and Creativity*. Springer, Berlin, 361–381.

Oliver Bown. 2014. Empirically grounding the evaluation of creative systems: Incorporating interaction design. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, Ljubljana, Slovenia, 112–119.

Kevin Burns. 2015. Computing the creativeness of amusing advertisements: A Bayesian model of Burma-Shave's muse. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 29, 01 (2015), 109–128.

Donald T. Campbell. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review* 67, 6 (1960), 380.

Linda Candy and Zafer Bilda. 2009. Understanding and evaluating creativity. In *Proceedings of the 7th ACM Conference on Creativity and Cognition*. ACM, 497–498.

Erin A. Carroll and Celine Latulipe. 2009. The creativity support index. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 4009–4014.

Heather Chan and Dan A. Ventura. 2008. Automatic composition of themed mood pieces. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*. Association for Computational Creativity, 19–115,28.

John Charnley, Alison Pease, and Simon Colton. 2012. On the notion of framing in computational creativity. In *Proceedings of the 3rd International Conference on Computational Creativity*. Association for Computational Creativity, 77–82.

Simon Colton. 2008. Creativity versus the perception of creativity in computational systems. In *AAAI Spring Symposium: Creative Intelligent Systems*. Association for the Advancement of Artificial Intelligence, 14–20.

Simon Colton. 2012. The painting fool: Stories from building an automated painter. In *Computers and Creativity*. Springer, Berlin, 3–38.

Simon Colton, A. Pease, and J. Charnley. 2011. Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*. Association for Computational Creativity, 90–95.

Simon Colton, Alison Pease, Joseph Corneli, Michael Cook, and Teresa Llano. 2014. Assessing progress in building autonomously creative systems. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, 137–145.

Simon Colton and Geraint A. Wiggins. 2012. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence*. IOS Press, 21–26.

Joseph Corneli, Anna Jordanous, Rosie Shepperd, Maria Teresa Llano, Joanna Misztal, Simon Colton, and Christian Guckelsberger. 2015. Computational poetry workshop: Making sense of work in progress. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 268–275.

D. H. Cropley and A. J. Cropley. 2005. Engineering creativity: A systems concept of functional creativity. In *Creativity Across Domains: Faces of the Muse*. Lawrence Erlbaum Associates, Inc., 169–185.

David H. Cropley, James C. Kaufman, and Arthur J. Cropley. 2008. Malevolent creativity: A functional model of creativity in terrorism and crime. *Creativity Research Journal* 20, 2 (2008), 105–115.

Mihaly Csikszentmihalyi. 1996. *Creativity: Flow and the Psychology of Discovery and Invention*. Harper Collins, New York.

Mihaly Csikszentmihalyi. 1999. Implications of a systems perspective for the study of creativity. In *Handbook of Creativity*. Cambridge University Press, Cambridge, UK, 313–338.

Joao M. Cunha, Joao Gonçalves, Pedro Martins, Penousal Machado, and Amílcar Cardoso. 2017. A pig, an angel and a cactus walk into a blender: A descriptive approach to visual blending. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 80–87.

Palle Dahlstedt. 2012. Between material and ideas: A process-based spatial model of artistic creativity. In *Computers and Creativity*. Springer, Berlin, 205–233.

Amitava Das and Björn Gambäck. 2014. Poetic machine: Computational creativity for automatic poetry generation in Bengali. In *Proceedings of the 5th International Conference on Computational Creativity (ICCC'14)*. Association for Computational Creativity, Ljubljana, Slovenia, 230–238.

Subrata Dasgupta. 2011. Contesting (Simonton's) blind variation, selective retention theory of creativity. *Creativity Research Journal* 23, 2 (2011), 166–182.

Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C. Neubauer. 2015. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts* 9, 1 (2015), 35.

Steve DiPaola, Graeme McCaig, Kristin Carlson, Sara Salevati, and Nathan Sorenson. 2013. Adaptation of an autonomous creative evolutionary system for real-world design application based on creative cognition. In *Proceedings of the4th International Conference on Computational Creativity*. Association for Computational Creativity, 40–47.

Alan Dorin and Kevin B. Korb. 2012. Creativity refined: Bypassing the gatekeepers of appropriateness and value. In *Computers and Creativity*. Springer, Berlin, 339–360.

Shlomo Dubnov, Kevin Burns, and Yasushi Kiyoki. 2016. Cross-cultural aesthetics: Analyses and experiments in verbal and visual arts. In *Cross-Cultural Multimedia Computing*. Springer, Switzerland, 21–41.

Ernest Edmonds, Lizzie Muller, and Matthew Connell. 2006. On creative engagement. *Visual Communication* 5, 3 (2006), 307–322.

Arne Eigenfeldt, Oliver Bown, Andrew R. Brown, and Toby Gifford. 2017. Distributed musical decision-making in an ensemble of musebots: Dramatic changes and endings. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 88–95.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 96–103.

Ahmed Elgammal and Babak Saleh. 2015. Quantifying creativity in art networks. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 39–46.

Frieda Fayena-Tawil, Aaron Kozbelt, and Lemonia Sitaras. 2011. Think global, act local: A protocol analysis comparison of artists' and nonartists' cognitions, metacognitions, and evaluations while drawing. *Psychology of Aesthetics, Creativity, and the Arts* 5, 2 (2011), 135.

Celso França, Luıs Fabrıcio W. Góes, Alvaro Amorim, Rodrigo Rocha, and Alysson Ribeiro da Silva. 2016. Regent-dependent creativity: A domain independent metric for the assessment of creative artifacts. In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, 68–76.

Ronald S. Friedman and Christa L. Taylor. 2014. Exploring emotional responses to computationally-created music. *Psychology of Aesthetics, Creativity, and the Arts* 8, 1 (2014), 87.

Philip Galanter. 2012. Computational aesthetic evaluation: Past and future. In *Computers and Creativity*. Springer, Berlin, 255–293.

Rodrigo García, Pablo Gervás, Raquel Hervás, Rafael Pérez, and Fernando ArÃmbula. 2006. A framework for the ER computational creativity model. In *Proceedings of MICAI 2006: Advances in Artificial Intelligence*. Springer, 70–80.

Berys Gaut. 2010. The philosophy of creativity. *Philosophy Compass* 5, 12 (2010), 1034–1046.

Pablo Gervás. 2002. Exploring quantitative evaluations of the creativity of automatic poets. In *Proceedings of the 15th European Conference on Artificial Intelligence Workshop on Creative Systems, Approaches to Creativity in Artificial Intelligence and Cognitive Science,* European Association for Artificial Intelligence, 8.

Pablo Gervás. 2010. Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 2nd Workshop on Computational Approaches to Linguistic Creativity*. Association for Computational Linguistics, 23–30.

Pablo Gervás. 2013. Computational modelling of poetry generation. In *Artificial Intelligence and Poetry Symposium, AISB Convention*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, Exeter University, UK, Article 3, 6 pages.

Vlad Petre Glăveanu. 2015. Creativity as a sociocultural act. *The Journal of Creative Behavior* 49, 3 (2015), 165–180.

Fernand Gobet and Herbert A. Simon. 1998. Expert chess memory: Revisiting the chunking hypothesis. *Memory* 6, 3 (1998), 225–255.

João Gonçalves, Pedro Martins, and Amílcar Cardoso. 2017. Blend city, blendville. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 112–119.

Kazjon Grace, Mary Lou Maher2 Maryam Mohseni, and Rafael Pérez y Pérez. 2017. Encouraging p-creative behaviour with computational curiosity. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 120–127.

Oskar Gross, Hannu Toivonen, Jukka M. Toivanen, and Alessandro Valitutti. 2012. Lexical creativity from word associations. In *Proceedings of the 7th International Conference on Knowledge, Information and Creativity Support Systems*. IEEE, 35–42.

Christian Guckelsberger, Christophe Salge, and Simon Colton. 2017. Addressing the "why?" in computational creativity: A non-anthropocentric, minimal model of intentional creative agency. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 128–135.

Sarah Harmon. 2015. FIGURE8: A novel system for generating and evaluating figurative language. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 71–77.

Paul Hekkert, Dirk Snelders, and Piet C. W. Wieringen. 2003. "Most advanced, yet acceptable": Typicality and novelty as joint predictors of aesthetic preference in industrial design. *British Journal of Psychology* 94, 1 (2003), 111–124.

Jiun-Jhy Her. 2014. An analytical framework for facilitating interactivity between participants and interactive artwork: Case studies in MRT stations. *Digital Creativity* 25, 2 (2014), 113–125.

I. Infantino, A. Augello, A. Manfré, G. Pilato, and F. Vella. 2016. ROBODANZA: Live performances of a creative dancing humanoid. In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, 388–395.

Arthur M. Jacobs. 2015. Towards a neurocognitive poetics model of literary reading. In *Cognitive Neuroscience of Natural Language Use*. Cambridge University Press, Cambridge, UK, 135–159.

D. A. Johner, D. Bedwell, C. Graham, W. Lemmon, O. Martinez, and A. K. Goel. 2015. Using human computation to acquire novel methods for addressing visual analogy problems on intelligence tests. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 23–30.

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

Anna Jordanous. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, 129–136.

Anna Jordanous. 2016a. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28, 2 (2016), 194–216.

Anna Jordanous. 2016b. The longer term value of creativity judgements in computational creativity. In *Proceedings of the AISB Symposium on Computational Creativity*. AISB, 16–23.

Anna Jordanous. 2016c. Personal communication on Twitter. (2016). Retrieved June 2016 from http://twitter.com/annajordanous/status/747766290648080384.

Anna Jordanous, Daniel Allington, and Byron Dueck. 2015. Measuring cultural value using social network analysis: A case study on valuing electronic musicians. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 110–117.

Anna Katerina Jordanous. 2012b. *Evaluating Computational Creativity: A Standardised Procedure for Evaluating Creative Systems and Its Application*. Ph.D. dissertation. University of Sussex.

Patrik N. Juslin, Laura S. Sakka, Gonçalo T. Barradas, and Simon Liljeström. 2016. No accounting for taste? Idiographic models of aesthetic judgment in music. *Psychology of Aesthetics, Creativity, and the Arts* 10, 2 (2016), 157.

Anna Kantosalo and Hannu Toivonen. 2016. Modes for creative human-computer collaboration: Alternating and task-divided co-creativity. In *Proceedings of the7th International Conference on Computational Creativity*. Association for Computational Creativity, 77–84.

Anna Aurora Kantosalo, Jukka Mikael Toivanen, Hannu Tauno Tapani Toivonen, and others. 2015. Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 276–283.

Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. In *Proceedings of the 14th International Conference on Advanced Learning Technologies*. IEEE, 508–512.

James C. Kaufman and John Baer. 2012. Beyond new and appropriate: Who decides what is creative? *Creativity Research Journal* 24, 1 (2012), 83–91.

James C. Kaufman, John Baer, Jason C. Cole, and Janel D. Sexton. 2008. A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal* 20, 2 (2008), 171–178.

James C. Kaufman and Ronald A. Beghetto. 2009. Beyond big and little: The four c model of creativity. *Review of General Psychology* 13, 1 (2009), 1.

Kyungil Kim, Jinhee Bae, Myung-Woo Nho, and Chang Hwan Lee. 2011. How do experts and novices differ? Relation versus attribute and thinking versus feeling in language use. *Psychology of Aesthetics, Creativity, and the Arts* 5, 4 (2011), 379.

Alexis Kirke and Eduardo Miranda. 2013. Emotional and multi-agent systems in computer-aided writing and poetry. In *Proceedings of the Artificial Intelligence and Poetry Symposium*. AISB, Article 4, 6 pages.

Maria E. Kronfeldner. 2010. Darwinian "blind" hypothesis formation revisited. *Synthese* 175, 2 (2010), 193–218.

Carolyn Lamb, Daniel G. Brown, and Charles L. A. Clarke. 2015. Human competence in creativity evaluation. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 102–109.

Jon O. Lauring, Matthew Pelowski, Michael Forster, Matthias Gondan, Maurice Ptito, and Ron Kupers. 2016. Well, if they like it… effects of social groups' ratings and price information on the appreciation of art. *Psychology of Aesthetics, Creativity, and the Arts* 10, 3 (2016), 344.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology* 95, 4 (2004), 489–508.

Helmut Leder, Gernot Gerger, David Brieber, and Norbert Schwarz. 2014. What makes an art expert? Emotion and evaluation in art appreciation. *Cognition and Emotion* 28, 6 (2014), 1137–1147.

Helmut Leder and Marcos Nadal. 2014. Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–Developments and challenges in empirical aesthetics. *British Journal of Psychology* 105, 4 (2014), 443–464.

John Lee, Ying Cheuk Hui, and Yin Hei Kong. 2016. Knowledge-rich, computer-assisted composition of chinese couplets. *Digital Scholarship in the Humanities* 31, 1 (2016), 152–163.

Joel Lehman and Kenneth O. Stanley. 2012. Beyond open-endedness: Quantifying impressiveness. In *Artificial Life*, Vol. 13. MIT Press, CambridgeS, 75–82.

Simo Linkola, Anna Kantosalo, Tomi Männistö, and Hannu Toivonen. 2017. Aspects of self-awareness: An anatomy of metacreative systems. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 189–196.

Simo Linkola, Tapio Takala, and Hannu Toivonen. 2016. Novelty-seeking multi-agent systems. In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, 1–8.

Maria Teresa Llano, Rose Hepworth, Simon Colton, Jeremy Gow, John Charnley, N. Lavrac, M. Znidaršic, Matic Perovšek, Mark Granroth-Wilding, and Stephen Clark. 2014. Baseline methods for automated fictional ideation. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, 211–219.

Róisın Loughran and Michael O'Neill. 2016. Generative music evaluation: Why do we limit to human? In *Proceedings of the 1st Conference on Computer Simulation of Musical Creativity*. University of Huddersfield, Huddersfield, UK, Article 10, 16 pages.

Róisın Loughran and Michael O'Neill. 2017. Application domains considered in computational creativity. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity.

Todd I. Lubart. 1999. 17 creativity across cultures. In *Handbook of Creativity*. Cambridge University Press, Cambridge, UK, 339–350.

Mary-Anne Mace and Tony Ward. 2002. Modeling the creative process: A grounded theory analysis of creativity in the domain of art making. *Creativity Research Journal* 14, 2 (2002), 179–192.

Jon McCormack and Mark d'Inverno. 2014. On the future of computers and creativity. In *Proceedings of the AISB Symposium on Computational Creativity*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, London, UK, 1–4.

Stephen McGregor, Matthew Purver, and Geraint Wiggins. 2016. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*. ACM SIGGEN, Edinburgh, UK, 51.

Marvin L. Minsky. 1982. Why people think computers can't. *AI Magazine* 3, 4 (1982), 3.

D. Moffat and M. Kelly. 2006. An investigation into people's bias against computational creativity in music composition. In *Proceedings of the 3rd Joint Workshop on Computational Creativity (ECAI'06)*.

Kristine Monteith, Bruce Brown, Dan Ventura, and Tony Martinez. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the 1st International Conference on Computational Creativity*. Association for Computational Creativity, 140–149.

Martin Mumford and Dan Ventura. 2015. The man behind the curtain: Overcoming skepticism about creative computing. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 1–7.

Frieder Nake. 2012. Construction and intuition: Creativity in early computer art. In *Computers and Creativity*. Springer, Berlin, Germany, 61–94.

Santiago Negrete-Yankelevich and Nora Morales-Zaragoza. 2014. The apprentice framework: Planning, assessing creativity. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, 280–283.

David Norton, Derrall Heath, and Dan Ventura. 2010. Establishing appreciation in a creative system. In *Proceedings of the 1st International Conference on Computational Creativity*. Association for Computational Creativity, 26–35.

David Norton, Derrall Heath, and Dan Ventura. 2011. An artistic dialogue with the artificial. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*. ACM, 31–40.

David Norton, Derrall Heath, and Dan Ventura. 2013. Finding creativity in an artificial artist. *The Journal of Creative Behavior* 47, 2 (2013), 106–124.

David Norton, Derrall Heath, and Dan Ventura. 2015. Accounting for bias in the evaluation of creative computational systems: An assessment of DARCI. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 31–38.

Ana-Maria Olteţeanu and Zoe Falomir. 2015. comRAT-C: A computational compound remote associates test solver based on language data and its comparison to human performance. *Pattern Recognition Letters* 67 (2015), 81–90.

Ana-Maria Olteţeanu and Zoe Falomir. 2016. Object replacement and object composition in a creative cognitive system: Towards a computational solver of the alternative uses test. *Cognitive Systems Research* 39 (2016), 15–32.

Philippe Pasquier, Adam Burnett, and James Maxwell. 2016. Investigating listener bias against musical metacreativity. In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, 42–51.

Marcus Pearce, David Meredith, and Geraint Wiggins. 2002. Motivations and methodologies for automation of the compositional process. *Musicae Scientiae* 6, 2 (2002), 119–147.

Marcus Pearce and Geraint Wiggins. 2001. Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, York, UK, 22–32.

Alison Pease and Simon Colton. 2011. 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*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, York, UK, 15–22.

Alison Pease, Daniel Winterstein, and Simon Colton. 2001. Evaluating machine creativity. In *Proceedings of the Workshop on Creative Systems, 4th International Conference on Case Based Reasoning*. ACM, 129–137.

Francisco C. Pereira, Mateus Mendes, P. Gervás, and Amılcar Cardoso. 2005. Experiments with assessment of creative systems: An application of Ritchie's criteria. In *Proceedings of the Workshop on Computational Creativity, 19th International Joint Conference on Artificial Intelligence*. ACM, 05.

Jonathan A. Plucker and Joseph S. Renzulli. 1999. Psychometric approaches to the study of human creativity. In *Handbook of Creativity*. Cambridge University Press, Cambridge, UK, 35–61.

Emma Policastro and Howard Gardner. 1999. From case studies to robust generalizations: An approach to the study of creativity. In *Handbook of Creativity*. Cambridge University Press, Cambridge, UK, 213–225.

Oscar Puerto and David Thue. 2017. A model of inter-musician communication for artificial musical intelligence. In *Proceedings of the 8th International Conference on Computational Creativity*. Association for Computational Creativity, 221–228.

Mel Rhodes. 1961. An analysis of creativity. *Phi Delta Kappan* 42, 7 (1961), 305–310.

Mark O. Riedl and R. Michael Young. 2006. Story planning as exploratory creativity: Techniques for expanding the narrative search space. *New Generation Computing* 24, 3 (2006), 303–323.

Graeme Ritchie. 2001. Assessing creativity. In *Proceedings of the AISB01 Symposium.*

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

Iván Guerrero Román and Rafael Pérez y Pérez. 2014. Social mexica: A computer model for social norms in narratives. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational creativity, 192–200.

Kimiko Ryokai, Noriko Misra, and Yoshinori Hara. 2015. Artistic distance: Body movements as launching points for art inquiry. In *Proceedings of the 33rd Annual ACM Conference on Extended Abstracts on Human Factors in Computing Systems*. ACM, 679–686.

Eugene Sadler-Smith. 2015. Wallas' four-stage model of the creative process: More than meets the eye? *Creativity Research Journal* 27, 4 (2015), 342–352.

Jonathan Sammartino and Stephen E. Palmer. 2012. Aesthetic issues in spatial composition: Representational fit and the role of semantic context. *Perception* 41, 12 (2012), 1434.

Rob Saunders, Petra Gemeinboeck, Adrian Lombard, Dan Bourke, and A. Baki Kocaballi. 2010. Curious whispers: An embodied artificial creative system. In *Proceedings of the 1rst International Conference on Computational Creativity*. Association for Computational Creativity, 100–109.

Oscar Schwartz and Benjamin Laird. 2015. bot or not. Retrieved on Aug. 9. 2015 from http://botpoet.com/.

John R. Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3, 03 (1980), 417–424.

Jennifer G. Sheridan, Alan Dix, Simon Lock, and Alice Bayliss. 2005. Understanding interaction in ubiquitous guerrilla performances in playful arenas. In *People and Computers XVIII—Design for Life: Proceedings of HCI 2004*. Springer, Bornemouth, UK, 3–17.

Hirohito Shibata and Koichi Hori. 2002. A system to support long-term creative thinking in daily life and its evaluation. In *Proceedings of the 4th Conference on Creativity & Cognition*. ACM, Loughborough, UK, 142–149.

Paul J. Silvia. 2005. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology* 9, 4 (2005), 342.

Dean Keith Simonton. 2011. Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology* 15, 2 (2011), 158.

Michael R. Smith, Ryan S. Hintze, and Dan Ventura. 2014. Nehovah: A neologism creator nomen ipsum. In *Proceedings of the 5th International Conference on Computational Creativity*. Association for Computational Creativity, 173–181.

Tim Smithers. 1997. Autonomy in robots and other agents. *Brain and Cognition* 34, 1 (1997), 88–106.

A. Tapscott, J. Gómez, C. León, J. Smailović, M. Zňidaršič, and P. Gervás. 2016. Empirical evidence of the limits of automatic assessment of fictional ideation. In *Proceedings of the 5th International Workshop on Computational Creativity, Concept Invention, and General Intelligence at ESSLLI*. Association for Logic, Language and Information, Bozen-Bolzano, Italy, 58–71.

Brandon Tearse, Peter Mawhorter, Michael Mateas, and Noah Wardrip-Fruin. 2011. Experimental results from a rational reconstruction of MINSTREL. In *Proceedings of the 2nd International Conference on Computational Creativity*. Association for Computational Creativity, 54–59.

Ellis Paul Torrance. 1968. *Torrance Tests of Creative Thinking*. Personnel Press, Incorporated, Princeton.

Patrick Tresset and Oliver Deussen. 2014. Artistically skilled embodied agents. In *Proceedings of the AISB Symposium on Computational Creativity*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour, London, UK, Article 9, 8 pages.

Tony Veale. 2013. Linguistic readymades and creative reuse. *Journal of Integrated Design and Process Science* 17, 4 (2013), 37–51.

Tony Veale. 2015. Game of tropes: Exploring the placebo effect in computational creativity. In *Proceedings of the 6th International Conference on Computational Creativity*. Association for Computational Creativity, 78–85.

Dan Ventura. 2016. Mere generation: Essential barometer or dated concept? In *Proceedings of the 7th International Conference on Computational Creativity*. Association for Computational Creativity, 17–24.

Dan A. Ventura. 2008. A reductio ad absurdum experiment in sufficiency for evaluating (computational) creative systems. In *Proceedings of the 5th International Joint Workshop on Computational Creativity*. Association for Computational Creativity, 11–19.

Graham Wallas. 1926. *The Art of Thought.* Jonathan Cape. London.

Annalu Waller, Rolf Black, David A. O'Mara, Helen Pain, Graeme Ritchie, and Ruli Manurung. 2009. Evaluating the STANDUP pun generating software with children with cerebral palsy. *ACM Transactions on Accessible Computing (TACCESS)* 1, 3 (2009), 16.

Thomas B. Ward, Steven M. Smith, and Ronald A. Finke. 1999. Creative cognition. In *Handbook of Creativity.* Cambridge University Press, Cambridge, UK, 189–212.

Robert W. Weisberg. 2015. On the usefulness of value in the definition of creativity. *Creativity Research Journal* 27, 2 (2015), 111–124.

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

Received December 2016; revised November 2017; accepted November 2017