Limitations from Assumptions in Generative Music Evaluation

Róisín Loughran and Michael O'Neill,

Natural Computing Research and Applications Group (NCRA),
University College Dublin,
Ireland
roisin.loughran@ucd.ie

Abstract. The merit of a given piece of music is difficult to evaluate objectively; the merit of a computational system that creates such a piece of music may be even more so. In this paper, we propose that there may be limitations resulting from assumptions made in the evaluation of autonomous compositional or creative systems. The paper offers a review of computational creativity, evolutionary compositional methods and current methods of evaluating creativity. We propose that there are potential limitations in the discussion and evaluation of generative systems from two standpoints. First, many systems only consider evaluating the final artefact produced by the system whereas computational creativity is defined as a behaviour exhibited by a system. Second, artefacts tend to be evaluated according to recognised human standards. We propose that while this may be a natural assumption, this focus on human-like or human-based preferences could be limiting the potential and generality of future music generating or creative-AI systems.

Keywords: Autonomous systems, creativity, evaluation, music generation.

1 Introduction

Whether or not computers can actually display creativity is a thorny subject, one that is unlikely to be resolved in the immediate or even near future. This is in part due to the prickly nature of the general understanding of creativity and all this word implies, before a computational emanation of it is even considered. This lack of understanding naturally leads to a difficulty in quantifying or enumerating what it means to be creative or to display creativity; there is a subjective nature to creativity that is very difficult to measure empirically. This difficulty in subjective measures has resulted in most computationally creative systems being evaluated using human opinion. This is understandable because comparing a computer's displayed creativity against that which is understood as being human creativity would seem to be the best (or only) way to circumvent this inherent subjectivity. On the other hand, Boden posited an interesting take on how true computer creativity could be recognised in the future (1998, p. 355): 'The ultimate vindication of AI-creativity would be a program that generated novel ideas which initially perplexed or even repelled us, but which was able to persuade us that they were indeed valuable.' This suggestion of recognising computer creativity retrospectively as something we could not appreciate (or were 'repelled' by) when first exposed to it implies that we must look further afield than our own human opinion for evaluation of computational creativity. If we adjudicate a creative artefact merely according to whether or not it is 'liked' by a human consensus, then Boden's above hypothesis will be impossible to realise.

This article examines the phenomenon of creativity, computational creativity and in particular musical computational creativity and the importance of evaluating it in a meaningful and sufficient manner. In recent years, we have witnessed remarkable progress in the field of machine learning and artificial intelligence. We have seen a program beat a human champion in chess (Campbell, Hoane & Hsu, 2002) and more recently one of the oldest human board games in the world, Go (Silver *et al.*, 2016). While this illustrates human-competitive levels for computer programs at logical tasks such as gaming, when it comes to more subjective, creative tasks such as musical composition, people can be less accepting of a computer's ability to match that of a human. Music is considered beautiful, aesthetic and above all personal – we each have our own taste in music that is ours to own with no need to defend. Can we expect autonomous programs to create aesthetic artefacts such as musical compositions that are comparable or indistinguishable from those created by humans? Is the only method of evaluating the creativity of a program to compare it against human creations using human opinion? While this may seem like the natural option, we propose that this is a limiting assumption – one that may hinder the development of computational creativity. In this article we discuss various aspects of musical computational creativity and consider if there are alternative manners in which to think about computational creativity – other than as a method of mimicking human creativity.

The following section considers the semantics of the word 'create' and how everyday use of the word and its variants can affect the meaning interpreted from it. Section 3 considers algorithmic compositional methods that are focussed on evolutionary techniques. The idea of conceptual space and transformations that can result in creativity is discussed in Section 4. The role of intelligence and how it relates to creativity and music is discussed in Section 5. Section 6 describes and discusses a number of methods that have been used to evaluate computationally creative systems in the past. A discussion of why we have this tendency towards human-comparisons and how this has changed in the definition of computational creativity is offered in Section 7. Finally, some conclusions are offered in Section 8.

2 Create, Creation, Creative, Creativity

Boden has stated that creativity is not magical but a feature of human intelligence (Boden, 2009, p. 23). Yet somehow, when we talk about creativity or whether or not someone is creative, it does translate into more than a simple ability to create. The specific use of the terms 'create', 'creation', 'creative' and 'creativity' does infer a different internal meaning when used colloquially. Although this is merely a grammatical or semantic difference, the implications of what is assumed are worth noting.

Ritchie discusses difficulties in implications from the words 'creative' and 'creativity', noting the lack of scientific rigour in the use of these words in ordinary discourse (Ritchie, 2006, p. 242). The words creative and creativity in relation to the process of creating raises ambiguities in the colloquial uses of such terms. Even dictionary definitions of the terms 'create' and 'creative' can vary, as discussed in (Jordanous, 2012, p. 254). The ability to create, to make something, does not immediately instil awe or wonder in us. We encourage and expect pre-school children to create drawings, models or stories as part of early development. We assume that we all posses this innate ability to be able to create or make a creation. Once we switch terminology to being creative, however, we somehow assume that this is a special ability, only afforded to a lucky few. In contrast to the simple creative ability we attribute to small children, adult creativity can often be used to infer a special talent or artistic ability. When considering creativity in absolute terms, or in terms of recognising creative ability in any autonomous system, the meaning of what is to be expected must be clear.

2.1 Types of Creativity

A creative idea must have novelty and value, but this can mean many things. An idea can refer to a physical artefact – a painting, composition, joke – or it can refer to a more abstract concept, theory or interpretation. The term 'value' can be interpreted as having many meanings; the idea could be beautiful, interesting, useful, more efficient, etc. Furthermore, there are two distinct variations to the term 'novel'. Ideas that are novel to the individual who generated it are considered Psychologically (P) Creative, whereas ideas that are novel to the world – ones that no one has considered before are said to be Historically (H) Creative (Boden, 2009, p. 24). By this reasoning, H-Creativity is a special case of P-Creativity. P-Creativity is the type of creativity we display in our everyday lives – which we expect from small children as discussed above. H-Creativity, on the other hand, results in the big discoveries – the famous symphonies and Nobel Prize discoveries. It is likely that the assumption that creativity mostly refers to, or even aspires to, H-Creative feats instils this idea of 'magical' creativity in us; creative accomplishments appear to be reserved for those talented few. The ability to be creative, however, is possessed by us all. While very few of us may display H-Creativity at any point in our lives, we display P-Creativity every time we make a joke, solve a problem or hum a tune.

There are three distinct types of creativity: combinational, explorational and transformational (Boden, 2004, p. 4). Combinational creativity combines familiar ideas resulting in a new unfamiliar idea or concept. An analogy is a form of combinational creativity that combines familiar concepts. Combinational creativity is the type of creativity that is most often used in studying experimental psychology. Exploratory creativity relies on the notion of a 'conceptual space'. This space is defined and constrained implicitly according to the domain being considered; it is the space within which a creative idea can be iteratively explored. Transformational creativity involves the most drastic alterations of all methods. In transformational creativity, the space within which one is searching is itself altered. This type of creativity offers the greatest opportunity for discovery or 'shock value', but it is also the most difficult to evaluate, as the transformations make meaningful interpretation or evaluation criteria very difficult to define. This idea of conceptual space is considered further in Section 4 below.

2.2 Computational Creativity

Computational Creativity is a subfield of Artificial Intelligence (AI) research that focuses on computational systems that undertake creative ideas. There have been a number of variations on the definition of computational creativity as the field has developed, but for this article, we will consider that given by Colton and Wiggins (Colton & Wiggins, 2012, p. 21): 'The philosophy, science and engineering of computational systems which, by taking on specific responsibilities, exhibit behaviours that unbiased observers would deem to be creative.' Thus computational creativity is defined in terms of being deemed creative – a term easy to discuss and describe (as above) but still difficult explicitly to define. This inherent difficulty in defining creativity in general is inevitably transferred to the domain of computational creativity. Such a difficulty leads to a further difficulty in evaluating any such creativity. As discussed in Section 1, there remains a strong tendency to evaluate such things using human opinion, but we would like to emphasise that the above definition makes no reference to human-like or human-competitive behaviour;[1] this definition explicitly states that it is an unbiased observer that must deem the behaviour to be creative. Thus, we again suggest that we must look further than human comparison in the evaluation of creativity.

Of the three types of creativity described above, combinational is the easiest for humans and yet the most difficult for an AI to achieve (Boden, 2009, p. 25). This type of creativity requires access to a vast range of ideas and concepts that a human naturally builds up over time but which must be made explicitly available to an AI. Nevertheless there have been a number of studies in humour that have looked at computational combinational creativity (Binsted, Pain & Ritchie, 1997; Manurung *et al.*, 2008; Valitutti & Veale, 2016). Using AI to model exploratory creativity requires high expertise and deep insights into the problem domain. Artists and musicians can spend years gaining expertise in their respective domains. Using a computational system to generate novel and valuable ideas requires close consideration of this knowledge. Yet exploratory systems have been developed in these areas of art (Colton, 2012) and music (Cope, 2004). Transformational creativity is the most difficult type of creativity to control, because it requires domain knowledge that must be maintained even when this domain is transformed. Boden has posited that evolutionary computational methods may be best suited to transformational creativity (Boden, 2009, p. 29). We discuss evolutionary methods applied to algorithmic composition in Section 3.

2.2 Algorithmic Composition

Algorithmic Composition (AC) can be considered a computationally creative task, but only if the compositions display true originality and creativity. Systems that merely mimic or adapt previously composed music would not, on the surface, appear to be

[1] Nevertheless, earlier definitions, such as that in Wiggins (2006), did make comparison to 'human'.

creative. In saying that, David Cope has stated that that creativity does not come from a vacuum, but synthesizes the work of others (Cope, 2005, p. 87). Cope's algorithmic compositional system *EMI* (Experiments in Musical Intelligence) was created to generate music in a given style and was trained on a corpus of existing music, initially a set of Bach chorales. He developed this system further into *Emily Howell*, an algorithmic composer who has released albums in her own style. Cope's definition of creativity is based on new connections between ideas not otherwise considered connected (p. 11). He warns against confusing creativity with novelty (p. 51) and instead focuses on recombinance (or rules acquisition) and allusion. He hypothesises that all composers in part combine ideas from other composers in their own work, hence he considers recombinance to be at the core of his computer models of musical creativity (p. 127).

The motivation for applying computation to musical tasks was examined and discussed in detail in Pearce, Meredith & Wiggins (2002), whereby they determined four distinct reasons, namely algorithmic composition, design of compositional tools, computational modelling of musical styles, and computational modelling of music cognition. Clearly there is more to be learned by applying algorithms to compositional tasks than merely creating computer music, although arguably algorithmic composition is still the most creative of these tasks. In discussing the motivations and evaluation of the compositional aim, however, they determine that 'researchers often fail to adopt suitable methodologies for the development and evaluation of composition programs and this, in turn, has compromised the practical or theoretical value of their research' (2002, p. 1). Thus a fundamental issue in applying computational methods to composition lies in the evaluation of the systems created.

3 Evolutionary Composition

The three types of creativity, introduced above, describe three ways in which computers can simulate creativity (Boden, 2004, p. 3):

- Combining novel ideas.
- Exploring the limits of conceptual space.
- Transforming established ideas that enable the emergence of unknown ideas.

Grammar-based evolutionary methods such as Grammatical Evolution (GE) (Brabazon, O'Neill & McGarraghy, 2015) offer an interesting parallel to such processes. The 'combination of ideas' concept can be likened to the crossover operator used in evolutionary systems; similarly, 'exploration' can be likened to the mutation operator. The use of grammars in GE can facilitate the third idea of 'transformation' listed above. Thus we propose that grammar-based evolutionary systems are particularly suitable for creative tasks such as melody writing. The creation of melodies offers a particularly difficult computational challenge, because there is no absolute correct answer; judging whether one melody is better than another is inherently a subjective matter. Systems based on Evolutionary Computation (EC) methods require the use of a fitness function – a user-defined function that can give a numerical assessment as to whether one solution is better than another. The design of this fitness function is hence very problematic for subjective tasks such as algorithmic composition. Often, this problem is addressed by using a human as a fitness function, using a set of known musical rules or comparing the music to a given style or genre. Each of these methods is based on the assumption that human-made music is best (and consequently is what is being searched for). But there is already an abundance of music being created (by humans) that follows such rules, with more being created every day. In looking at algorithmic composition as a computational problem, we are given an opportunity to consider it from a different angle. Assuming that the music created by machines must automatically be judged in human terms is an assumption that has the potential to limit the capabilities of any computationally creative system (Loughran & O'Neill, 2016a).

EC methods are fundamentally based on Darwin's evolutionary theory of 'survival of the fittest'. A population of random solutions to a given problem is created and each solution is assigned a fitness according to how well it solves that problem. The solutions are then selected for survival and reproduction into the next generation based on this fitness. As this process is repeated, the overall population of solutions is improved and the best in the final population can be chosen as the solution to the given problem. In applying EC to composition, the conceptual space is defined by the representation, musical rules or grammars used. Each individual in the population is a melody or part of a melody. The representation of music, fitness function and manner in which the results are interpreted or combined into music are all design considerations for the experimental programmer. The following discussion introduces a number of experiments that used EC methods for compositional tasks.

EC methods were developed using problems that had a specific optimal solution, such as symbolic regression and the artificial ant trail. In developing these systems for aesthetic purposes, we should perhaps look at a broader way of using and interpreting them. These are tools for composers to use, and as tools they can be utilised as seen fit. Miranda examined three distinct approaches to using evolutionary methods in music: the engineering approach uses EC techniques in the field of sound synthesis; the creative approach uses EC in compositions; and the musicological approach searches for the origins of music by means of computer simulations (Miranda, 2004). An overview of earlier studies in EC for musical composition is offered in Burton & Vladimirova (1999), determining that Genetic Programming (GP) (Koza, 1992) methods perform better than those that use Genetic Algorithms (GA) (Goldberg & Holland, 1988). This may be unsurprising because GP methods use a tree-based structure whereas GAs are limited to a linear string in their representation. Hence, GP can represent more complex representations and operations – something that would be very useful in representing music.

GenJam (Biles, 1994) used a GA to evolve jazz solos, building solos from pre-generated MIDI sequences that were judged by a user to determine the fitness measure. The system has been modified and developed into a real-time, MIDI-based, interactive improvisation and performance system that regularly performs in mainstream venues (Biles, 2013). *VoxPopuli* is an interactive compositional tool that uses evolutionary methods in real-time algorithmic music composition using notes and chords (Moroni, Manzolli, Von Zuben & Gudwin, 2000). Dahlstedt has discussed how we may use EC as the basis of a wide range of tools but that in doing so we may have to relinquish some level of control (2007, 2009). More recently, adapted GAs have been used with local search methods to investigate human virtuosity in composing with unfigured bass (Munoz, Cadenas, Ong & Acampora, 2016), with a grammar to augment live coding in creating music with *Tidal* (Hickinbotham & Stepney, 2016), and with non-dominated sorting in a multi-component generative music system that could generate chords, melodies and an accompaniment with two feasible-infeasible populations (Scirea, Togelius, Eklund & Risi, 2016).

Evolutionary processes work well in aesthetic tasks such as music composition because they are generally non-deterministic. The evolution of a population offers so much scope and possibility that it is reminiscent of the music creation process – a solution is not linearly determined but instead emerges from a fluid, incremental process. As introduced above, the biggest issue in using EC for aesthetic purposes is in the design of the fitness measure. Individual solutions (compositions in the case of AC) can only survive on to the next generation if they are judged worthy according to a predetermined fitness measure designed by the programmer. Thus the problem becomes: how do we measure the musical fitness of the individual?

3.1 Measuring Fitness

The most obvious approach to developing an aesthetic-judgment-based fitness measure is to use a human as the fitness function. Such systems are referred to as Interactive EC (IEC). In these experiments, a human user must rate each individual in every given generation. The survival of that individual is then dependent on the value given by the user. These systems are very well suited to design and creative tasks because they remove the need to automate a subjective judgment. A number of systems have used IEC to successfully create melodies (Biles, 1994; Moroni *et al.*, 2000; Reddin, McDermott & O'Neill, 2009; Shao, McDermott, O'Neill & Brabazon, 2010). The biggest drawback with interactive methods is that they create a bottleneck, particularly in musical tasks. For the analysis of visual art, whereby the user can observe a number of creations concurrently, the fitness can be measured very quickly. For musical tasks, however, users need to listen to musical excerpts successively, rendering these methods very expensive. For IEC experiments in algorithmic composition, the experiments must be designed so that the user only has to listen to and adjudicate a small number of compositions before fatigue or boredom sets in. Every time an experiment is run a new set of listening tests (possibly with a new set of listeners) must be set up. This makes it very cumbersome to re-run experiments and so IEC experiments must be very carefully prepared. For this reason it is simpler and less costly to develop an automatic fitness function.

In some studies, the initial population only contains individuals that are already of high quality. Because of this, individuals can be randomly selected (regardless of fitness) for reproduction (Waschka II, 2007) or the entire population can be used in creating the composition (Eigenfeldt & Pasquier, 2012; Loughran, McDermott & O'Neill, 2016). The idea of a random fitness function is alien to EC programmers because it is nonsensical to evolve a population without any fitness measure. If the system uses *a priori* musical knowledge to ensure the entire population is of high fitness, then the search space is confined so that the evolutionary process can be used to traverse the space safely. This may not be considered a proper use of EC – but it can make good music.

The use of a traditional, autonomous measure of fitness may be more economical than IEC and make more sense than random selection, but such a measure is not easy to define. An overview of the most prevalent measures and ideas used to examine and evaluate melodies is given in de Freitas, Guimaraes & Barbosa (2012). They discuss ten attributes used in the evaluation of melodies based on pitch and rhythm measurements, concluding that previous approaches to formalise a fitness function for melodies have not comprehensively incorporated all measures. Nevertheless, many studies have used various types of autonomous fitness functions to drive EC systems to create music (Todd & Werner, 1999; Dahlstedt, 2007; Loughran, McDermott & O'Neill, 2015b, 2015a; Munoz *et al*., 2016).

System-Based Fitness A notable study demonstrated that in computationally creative evolutionary systems, it is only important that the fitness measure chosen need be defensible; what makes one creative item better than another may not be what a human would choose, but it must be a sensible, defensible and reproducible choice by the computer program. In other words there must be a logical and explainable method in assigning fitness measures. This was investigated using the idea of a preference function by measuring qualities such as specificity, transitivity and reflexivity to determine the choice of a system in a number of subjective tasks (Cook & Colton, 2015). Such a measure may not agree with what a human may choose as the best but, most importantly, it agrees with itself. This preference function chooses one item over another due to a logical system of comparing between items and determining a decisive preference. A related idea was proposed for a compositional GE system that based fitness on the concept of conforming to the popular opinion of the population (Loughran & O'Neill, 2016b). In this system a population of 'critics' were evolved on a corpus of melodies according to how well each individual critic agreed with the ranking of the melodies by the entire population. This best critic was then used as a fitness function to create a new melody that replaced one of the original melodies in the corpus and the cycle was repeated. This resulted in a complex adaptive system that was self-referential and autonomous once it had been initialised. This system was generalised from a ranking-based system to a cluster-based system in Loughran & O'Neill (2017). The purpose of the development of such systems is to remove any human-defined measures of aesthetic fitness, enabling a compositional system to be autonomous and unbiased from human influence.

3.3 What's the Objective?

The above argument only considers EC applications but other Machine Learning (ML) music creation systems suffer from the same dilemma. Any supervised ML algorithm needs an error function – a target it must aim towards. Backpropagation, used in Artificial Neural Networks such as the Multi-layered Perceptron, requires a mean-squared error, which requires a target. Similarly any other supervised ML algorithm needs an error function – a target it must try to approach or optimise towards.

Such targets are, however, completely misaligned with the human method of composing. Human composers do not start with a target composition and iterate towards that. Students of academic music may be given assignments in which they must conform to a set of theoretical rules or emulate a given composer's style – but this is not where great compositions come from. Is the purpose of applying AI to music to produce a bunch of mediocre students or to create new, genuinely good and novel music?

One problem with traditional fitness functions is that they result in good or bad results, leading to a scale of 'goodness' depending on how close an individual is to a specified objective. Some AI researchers would propose that using a pre-specified objective is not necessarily a good idea when searching a space to solve a problem. This theory suggests that searching for novelty is a better method in looking for a great solution, in that the optimal solution can often be found when looking for a different solution or when searching for no particular solution at all (Lehman & Stanley, 2010; Stanley & Lehman, 2015). Such a theory fits very well in searching any creative space. A musician does not know what music they are trying to create when they start; they work through ideas, changing their process and hence their output as they observe what they are creating. We propose that for any automated machine-learning system to be truly creative there cannot be a pre-defined objective; the fitness function should be a measure of the progress of the system.

In recent years, the field of computational creativity has embraced this idea that creating an artefact means more than outputting a number. The context within which a creative product is judged, including background information and the feeling it evokes in the creator, is defined as Framing (Charnley, Pease & Colton, 2012). Such a concept reveals that there is more to computational creativity than the output, and that intent, motivation and aspects of the creative or computational process all contribute to the overall result. Similarly, a Computational Creativity Theory (CCT) has been proposed to provide a computationally detailed description of how creation could be generated and the impact it can have (Colton, Pease & Charnley, 2011). These studies demonstrate that there is more to measuring the progress of a creative system than merely taking a numerical measure of error, target or fitness.

3.4 Fitness *versus* Evaluation

In the case of using EC techniques for compositional tasks we must be very clear on the distinction between fitness measure and evaluation. The fitness is the continuous measure taken from individuals within the population that drives the evolution of the composition. Evaluation in this sense refers to the measure of the performance of the system as a whole – how successful the given system is at composing a piece of music. In creative tasks such as music creation, this results in a distinct disjunction between fitness measurement and the perceived quality of the output – one that is not present in more traditional, empirical uses of EC. We highlighted EC applications to music creation in this section because this fitness measure plays a crucial role although many other types of machine learning methods have been applied to the task of music composition (Fernandez & Vico, 2013). Regardless of the type of algorithm used, with any optimisation or error-based functionality, some metric of the aesthetic progress of the melody must be given throughout the composition process. This is not the same as evaluation, however. Evaluation involves measuring the overall success of the system either from the process involved or the final result produced, depending on the aim of the system.

4 Conceptual Space

The previous section drew parallels between evolutionary computation and computational creativity in terms of combinational, explorational and transformational creativity and the workings of evolutionary computation. While the terminologies used do offer an interesting conceptual analogy, a direct comparison is overly simplistic in regards to the space in which computational creativity is studied: the conceptual space. This conceptual space can be thought of as the abstract location of the artefacts produced by the creative system. As defined by Boden (2004, p. 4): 'Conceptual spaces are structured styles of thought... any disciplined way of thinking that is familiar to (and valued by) a certain social group.' Depending on the constraints of the given problem domain, this space can be sparsely or densely populated. In any given conceptual space, many thoughts may be valid or possible, but only some of them will actually be thought. Some thoughts may be obvious and natural and are reached without any effort or conscious deliberation. Others involve a deeper traversing of this space, to find the links to ideas not immediately obvious to us.

Both exploratory and transformational creativity are linked to this idea of the conceptual space. Exploratory creativity searches and traverses this space in generating novel ideas, whereas transformational creativity transforms a dimension of the space so that new ideas can be formed that would not have previously adhered to the space. Depending on the degree of transformation or the degree of exploration, these two forms of creativity can be seen to be operationally quite similar. Exploration of the space can be seen as a small 'tweaking' of some defined constraint that amounts to a minor transformation. The distinction between tweaking and transforming can be specific to the domain, but it is dependent on how well defined the concept space is (Boden, 1998, p. 348).

We are not aware of any attempt to define how many dimensions may be in a concept space, however in idea management systems an idea space has been suggested which was reduced in dimensionality (Spencer, 2012). This study proposed that by using feature-based Jaccard-Tanimoto similarity, this 'idea space' was consistently about 14-dimensional, regardless of the origin or specifics of the ideas. Although this result may appear over-simplified, the proposal to reduce such a space is interesting and may warrant further consideration.

4.1 Transformational creativity

Boden has posited that many big scientific discoveries involved some form of transformation, but many people believe computers could not achieve this type of transformational creativity (Boden, 2009, p. 29). The idea that transformational creativity results in the highest levels of creativity was formulated by Ritchie as a hypothesis for experimental testing (Ritchie, 2006). This study reconsiders some fundamental assumptions on computational creativity in a formal and informal sense.

Informally, he considers that a creative action takes place in a society of individuals resulting in an artefact. Within the society there is a small set of medium types and genres. An artefact belongs to a medium type, which merely indicates a raw data type of an artefact – a string of characters, etc. A genre is a culturally defined type of an artefact. The medium type of an artefact is trivial to decide, but which genre it belongs to may be made subjectively. In this discussion, Ritchie reconsiders a more abstract requirement for the conceptual space. In doing so, he considers a number of functions that a space must fulfil in order to support an analysis of creativity (p. 250). He states that whatever a space is, formally it must be something that can be abstracted from a set of artefacts. Thus, the given space must be able to hold all artefacts that exist within it. He considers a number of options for a formal model, but find no obvious formal distinction between minor and major changes or indeed whether a change would amount to the altering of a boundary of a space or multiple spaces. He notes that a transformation cannot be sufficient criteria for high creativity – merely a necessary one, while pointing out that this has not been verified in human creativity (p. 259). To test this hypothesis, he states that a precise formal model (of one of the types discussed in the paper) must be developed and that the space, the space induction and the transformation must be defined. He argues that while this is not trivial, if it cannot be done then empirical testing of the hypothesis would be impossible (p. 260). While he states that such an approach may not be the only option, anyone trying to assume transformational creativity is superior to other forms should offer some similar or comparable analysis (p. 263).

4.2 The Creative Step

Creativity is a step-wise process. Creativity cannot exist in a vacuum, nor can it just appear, but instead it must be reached through combination, exploration or transformation. Thus we propose that there must always be a ‘Creative Step’ – a movement from one idea to the next that results in the emergence of a sufficiently novel yet interesting idea. The size or extent of this step is critical in the recognition and perception of creativity. If this step is too wide, the creativity is lost as being random or nonsensical, but if it is too narrow it is too trivial to actually be creative. This is evident in artefacts as well as ideas. A Pollock painting would surely have been ridiculed in the 18th Century, but through gradual explorations and transformations within the artistic conceptual space it now may be revered as great work. The painting (artefact) could physically exist at either time, but it is only the changing appreciation of artistic works over time that can result in this painting being perceived as creative.

5 Musicality, Creativity and Intelligence

This article considers the implications of assumptions made in evaluating computationally created music or, more generally, in asking if an AI can be musically creative. The relationship between intelligence, creativity and music is clearly both complicated and yet highly important to establish in considering these ideas. One would naturally assume that the act of displaying creativity inherently displays intelligence. Indeed, Boden has described creativity as 'a feature of human intelligence in general' (Boden, 1998, p. 347). One would also assume that displaying musicality naturally displays creativity; it is fair to assume that if someone was to write an acceptably pleasing piece of music that this person would be considered creative. If the transitive property was to hold in this space, then we could state that in displaying musicality one is inherently displaying intelligence. While the level of intelligence that a display of musicality (or creativity) actually indicates is certainly debatable, a conflicting example of a system or person completely lacking in intelligence producing something musical does appear to be implausible. The converse of this is not true, however; there are many creative people that are not musical, just as there are many intelligent people that are not musical and would not claim to be creative.

In the non-human or machine context, Artificial Intelligence became a computing priority long before computational (or artificial) creativity became a topic of interest. Hence we know that there are many extant AI systems whose priority was not to display or consider any creativity. But, as per the argument above, does a system that displays musicality automatically display intelligence? If we again assume that musicality implies creativity and alter Boden's above description of creativity to state 'a feature of intelligence in general' rather than 'a feature of human intelligence', then we can state it does.

A more in-depth discussion on the relationship between music, intelligence and artificiality is offered in Marsden (2000). In this study, that appeared before many of the formal papers on computational creativity, the discipline of Music-AI is studied by considering two possibilities of machines: the idea of computers imitating human behaviour and also performing musical tasks. In this study the distinction between machines and other artificial objects is defined by their behaviour. From the point of view of information technology it makes the point that we value machines for what they can do, not what they are; computers were designed to have unconstrained behaviour, to be the universal programming machine. In discussing the history of Music-AI Marsden states that one characteristic of an intelligent being is that it can learn, not just from explicit teaching but that it can learn spontaneously. In a philosophical discussion on the definition of music and how an artificial system may be defined to be musical he states: '... if any system is to be musical it must make reference to human behaviour, and to that extent any musical system must involve artificial intelligence' (2000, p. 21). Thus Marsden states that when determining the musicality of a system, there is no obvious boundary to be drawn between considering human behaviour that is not intelligent and considering (non-human) behaviour that is intelligent. In this sense 'musicality' and 'intelligence' are very much intertwined, and very much dependent on emanating human-like behaviour. With the ever increasing computational power of machines, often what is expected of them is not equal-to-human but superhuman abilities; computers can process more data, faster than any human ever could. Marsden proposes that the real goal of an AI is for it to perform in a human-like manner in some respects and a non-human manner in others; but again this leads to questionable boundaries as to what constitutes 'human-like' and when human-like should be prioritised over non-human-like. Throughout the paper, Marsden considers three types of definitions of intelligence: behaving human-like, exhibiting spontaneous learning and responding to the surrounding environment. From this third definition, he states that one must consider the possibility that AI is not necessarily a copy of human intelligence. He proposes that this would offer interesting and productive (or valuable) and novel approaches which would be very interesting to musicians and in particular for the musical task of composing.

An interesting angle in the above study is its comparison between AI and intelligence in humans and animals. A thought provoking, if at times whimsical, comparison between an artificial mind and the mind of a dog is discussed at length in McFarland (2009).

6 Evaluating Creativity

The idea of evaluating creativity in terms of human opinion is nearly always assumed but rarely justified. One explicit justification for this is offered in Ritchie (2006). He defends this standpoint on two grounds. Firstly, he states that humans have used the term first and so this is the meaning that is well established. Secondly, Ritchie argues that to measure machine creativity in terms of mere machine performance could lead to the danger of circularity in claims about the nature of this process. The first of these arguments appears weak; justifying using humans merely because we used the term 'creative' first is not a very strong point. The second point on the danger of circularity due to the lack of clarity of the definition of the term 'creative' is a much stronger argument. The difficulty in defining creativity naturally leads to a resultant difficulty in evaluating whether or not a computational system is creative. This has led to a number of authors undertaking self-evaluation, minimal evaluation or no evaluations at all on their systems. The lack of evaluation in CC systems has been noted throughout the development of the field (Boden, 1998; Cardoso, Veale & Wiggins, 2009; Jordanous, 2011). Such studies highlight the need for a clear definition of what can be considered creative.

Ritchie was one of the first to propose a set of formal empirical criteria for creativity. He originally proposed a set of 14 criteria (Ritchie, 2001), which was extended to 18 (Ritchie, 2007) as a framework describing the design and implementation of a creative system. These criteria aim to judge the two main aspects of creativity – namely typicality (or, in contrast, novelty) and value or quality. The individual criteria are weighted in various ways to determine the quality and typicality of the produced output in comparison to what the system is expected to produce. Colton designed a framework entitled the Creative Tripod to determine if a system is creative, or if it merely has the perception of being creative (Colton, 2008). The Tripod framework describes a system as creative if it exhibits three elements: skill, appreciation and imagination. Furthermore, the framework states that there are three involved parties that may be perceived as contributing to this creativity, namely the programmer, the computer and the consumer. For creativity to be experienced, all three elements must be exhibited by at least one of these three parties. This is an extremely important step in the description and definition of creativity because it can separate the idea of creativity from the human user. If a programmer shows no creativity but the program she creates does, then creativity is present. A framework for evaluating genre-specific compositions was proposed in Pearce & Wiggins (2001). In this work they describe a framework that examines each phase of a generative music system culminating in a discrimination test. This evaluation was performed by human subjects by asking them how well the generated music conformed to a pre-specified genre. Pearce and Wiggins use a subsequent study to evaluate melodies with a learning-based perceptual model of music listening (Pearce & Wiggins, 2007). This study involved using a number of experienced human observers to judge the output of three computational methods of creating chorales, and then statistically analysing their judgements in order to help develop towards an autonomous creative system. This work proposes an excellent study in modelling a computational system on measured cognitive behaviour, but they acknowledge that their results suggest that these compositional tasks still present significant challenges in modelling cognitive processes.

A further discussion of various methods of evaluation applied to musically creative systems is given in Ariza (2009). He discusses the Musical Directive Toy Test (MDtT), whereby an interrogator, using a computer interface, gives a musical directive to two composers, one human and one machine. The given directive may be a style or abstract instruction and the interrogator must decide which output is from the human. A similar Musical Output Toy Test (MOtT) is described whereby two composers (again one human one machine) produce a piece of music that may be related in terms of style or instrumentation but are created without specific directive. Again the goal is to convince the interrogator that they are the human composer. Ariza compares the application of these tests in numerous studies but note that these tests, unlike the traditional Turing Test, do not rely on or require natural language, and that the decision made by the interrogator may rely as much on preference or subjective judgments as on logic. He proposes that the continued use of such tests does more to 'investigate the limits of musical judgement than the innovation of generative music systems' (Ariza, 2000, p. 57).

Currently, the most highly recommended system for evaluating creative systems is the Standardised Procedure for Evaluating Creative Based Systems (SPECS) (Jordanous, 2012; Jordanous, 2013). This work performed an initial survey of evaluative practice in contemporary (from 2007–2010) computational creative systems and papers. Jordanous found that evaluation of computational creativity was not being performed in a systematic or rigorous manner. She observed that these results indicate computational systems are being presented as 'creative systems' without justification of this creativity; the term 'creative' has become another descriptor of the system, rather than the focus of such systems. Furthermore, the survey in Jordanous (2012) drew attention to a lack of clarity as to what should be involved in evaluating a creative system – what interpretation of creativity should be used, who should perform evaluation and when, etc. Jordanous identified a lack of universally accepted and comprehensive definition as to what it means to be creative as a major complication in developing a standard or consistent method of evaluation. From the linguistic analysis performed on a review of literature over 60 years of creativity research, Jordanous identified 14 distinct components that act as building blocks for creativity. These components were used in developing a set of Evaluation Guidelines (Jordanous, 2011), involving three distinct steps to clarify what is being evaluated and then performing tests according to that clarification. For any developed creative system one must:

Step 1: Identify a definition of creativity that your system should satisfy to be considered creative

Step 2: Using this definition, clearly state the standards you use to evaluate the creativity of your system.
Step 3: Test your creative system against the standards stated in step 2 and report the results (Jordanous 2012, p. 259).

These guidelines were expanded into methodological steps that encompass the SPECS methodology (Jordanous, 2012). The SPECS framework has become a suggested standard for evaluation of creative systems (see, for example, the guidelines for this journal (JCMS, 2017)).

6.1 Artefact *versus* Behaviour

Unfortunately, a number of previous evaluation methods only evaluate the output artefact created by the system and do not consider the process or behaviour of the system itself. In fact, one of the Open Problems in Evolutionary Music and Art (McCormack, 2005, p. 434) states that it is important to create evolutionary art (or music) recognised by humans for it's artistic contribution as opposed to technical fascination. This is in direct contrast to the definition of computational creativity given in Colton & Wiggins (2012, p. 21) which is based on 'exhibited behaviour' of the system – and is not defined in terms of the output, or in terms of human opinion. This is an important distinction to be aware of in this stage of developing autonomous creative systems. If a system composes music it is very interesting to hear what kind of music it composes, but if it is the system's ability to create being evaluated, it is imperative to look further than the output in making this evaluation. The relevance of this distinction is dependent on the focus of one's research. Pearce and Wiggins specify two ways in which machine composers may be evaluated: in terms of the music they compose and in terms of the manner in which they compose (2001). There are many music generative systems and human-interactive systems whose purpose is to create music while other studies are more focussed on the academic exploration of autonomous musicality or creativity. Music systems focussed on 'mere' generation are defended in Eigenfeldt, Bown, Brown & Gifford (2016), highlighting that much music innovation has been achieved in Musical Metacreation (MuMe, 2017) from generative systems focussed on human-interactive co-creativity. As such, any given music generative system lies somewhere on a spectrum between pure generation (the artefact is most important) and pure computational creativity (the behaviour is most important). Meaningful and relevant evaluation of any system is dependent on where the system lies within such a spectrum.

This is not a new distinction to make. John Cage's *4'33"* is undeniably recognised as a musical work, but this is the classic example of appreciating the method or concept used over the output. Similarly, the serial works of Schoenberg and many Musique Concréte works are as focussed on the way in which the sounds within a piece are made as the final output. Such works also caused controversy (certainly outside an academic music audience) in their time, but they have stood the test of time and are recognised as landmarks in musical history. Such precedence should leave us open to more generalised methods of evaluations beyond whether or not people 'like' it.

6.2 The Lovelace Test

Often known as a founder of computer programming, Ada Lovelace had some remarkable insights into the possibilities that computer programming could offer. In the 1840s, Lovelace saw a capability in Charles Babbage's recently proposed Analytical Engine far greater than that of mere numerical manipulations. She saw that such machines could in time be used to represent art and music, but she maintained that these machines would never be able to create, as creation requires originating something. Her objections have been paraphrased by Bringsjord, Bello & Ferrucci (2003, p.4): 'But computers originate nothing; they merely do that which we order them, via programs, to do'. Her considerations on this topic were remarkable in relation to such a new theoretical invention at the time. In her writings, Lovelace posed a number of questions in this regard, which have been distinguished by Boden into the four 'Lovelace Questions' (Boden, 2004, p. 16). These questions ask:

- Can computational ideas help us understand human creativity?
- Can computers ever do things that appear creative?
- Can computers ever recognise creativity?
- Can computers ever *really* be creative?

Most people would agree that the first two questions have been answered (with a resounding 'yes'). The third may offer more argument, but it is the fourth question that causes the most bother to people. Bringsjord *et al.* (2003) consider that Lovelace posed these questions as an objection to the idea that computers could actually be creative. They note her objection that creation requires the origination of something whereas computers are not capable of originating anything. They subsequently developed the aptly named Lovelace Test (LT) for creativity. This test involves an artificial agent, A, its output, o, and its human architect, H. Simply put, the test is passed if H cannot explain how A produced o. While this may seem like simple criterion, it is actually extremely difficult to pass. This test requires that the algorithm written by the programmer must produce an output that the programmer, or another agent with the programmer's expertise, cannot explain. On the surface it may seem like many AC systems would quickly pass this. EC compositional systems, for example (see Section 3), having a non-deterministic nature, can produce output not predictable by the programmer. Not predictable is not the same as not explainable, however. The programmer can explain the representation, fitness measures or grammars used in such systems, thus explaining the process of how the music is produced. For the LT to be passed, the output must be truly surprising and unexplainable to the programmer.

The LT is much more difficult than other TT-style tests because it is the programmer, the one person who understands the workings of the algorithm more than anyone else, that acts as the interrogator of the system. In a sense the program must fool or trick its own creator for it to be deemed successful. If the programmer made a mistake, and suddenly could not remotely explain the output of her own system, would this be allowed to pass the LT? We would assume not, since a mistake implies randomness (on the programmer's part) and randomness is not equivalent to creativity. However, if the seemingly random human mistake led to a genuine creative streak, shouldn't this satisfy the specified criterion to pass the LT? Often our own most creative successes are attributed to a moment of inspiration. Could this not be seen as a 'mistake' in the mind that we cannot explain? If we can accept the results of our own random mistakes as creative, why does it need so much more explanation in the programs we create and, paradoxically, why is it that once we can explain it, it no longer can be claimed as creative?

The LT can be seen as an attempt to satisfy the fourth Lovelace Question posed above, and therein lies the difficulty. To be really creative is something that many humans feel they can only aspire to. The difficulties inferred by our colloquial use of the term 'creative' were discussed in Section 2. But creativity is not magic; it is not an elite quality only to be found in a lucky few, but an ability possessed by us all. The LT may be doomed to be impassable – by definition, if the programmer understands their own code, they can always offer some explanation as to the output that is produced. As algorithms become more complex, however, involving domain transformations, stochastic, statistical and non-deterministic measures, then surely this explanation will become a more abstract way of explaining how the output came about, rather than an exact explanation of how A produced o. Human artists are not held up to such scrutiny as to how they create a work of art. Critics may examine an artist through their teachers, mentors and influences, determining their reasoning for a given style according to what they have learned along their career path. This explanation of influences or learning does not negate the resultant creativity of a human artist. Why then should such an explanation automatically negate the creativity of an algorithm?

7 Discussion

Creativity must involve a display of reason and intent. Random acts that result in seemingly creative artefacts cannot be perceived as being creative. The current definition of computational creativity given above refers to systems that 'exhibit behaviour'; it does not in fact refer to the artefact produced. When evaluating a creative system it is vital to bear this in mind and not merely judge the system on the final output produced. The need for evaluating creative systems was discussed in the previous section but, while we do not dispute this, we want to mention studies that are purely focussed on the system rather than the output. For studies that are focussed on the method behind a system – for example, the architecture, level of autonomy or even an underlying concept – evaluation in the sense proposed may not be as important as it is for other more artefact-focussed systems. For such studies, is not submitting work for peer-review to suitable conferences or journals in itself a form of evaluation of the validity of the method or reasoning behind such a system?

7.1 Non-human Creativity

We have a tendency to anthropomorphise behaviours typically associated as being specifically human when we see such behaviour exhibited by non-human systems. For example, many dog owners consider their canine companions to have comprehension or understanding beyond what is empirically evident. A quick online search would offer a multitude of videos of dogs 'singing' along to music, humans singing, or to other dogs howling. Certain dogs may howl when they hear a musical instrument playing or a baby crying, but to say that this is singing along is attributing too much understanding and intent to an observed behaviour. Assuming animals have an aesthetic appreciation or enjoyment of music is ascribing a set of values that we possess onto a being that may not have the same set of values. No doubt the dog enjoys howling along (assuming she does it on her own accord), but this does not mean she appreciates music in the same sense as us. This concept of attributing a set of human values onto a non-human system or animal may seem natural, but it is questionable from a philosophical standpoint (McFarland, 2009). Such a philosophical argument is equally valid for an AI system. No AI system has yet been developed that exhibits intelligence to the level of that of a dog, yet we automatically assume that it will have the capabilities to generate or appreciate music in the same way as we do. Is that not again assuming too much for systems that are still in development?

7.2 The Human Comparison

In the development of the field of computational creativity, authors have defined and described creativity in terms of a 'human' ability to various extents. This has often been an implicit suggestion within the explanation of ideas or proposals. Boden has described creativity as something not magical but as an 'aspect of normal human intelligence' (Boden, 2009, p. 24). In Marsden's discussion on intelligence, music and artificiality he discusses the 'intention to perform in a human-like fashion' as one of the two major topics of the paper (Marsden, 2000, p. 16). Ritchie justifies alluding to human-creativity when considering more general (non-human or machine) creativity for two reasons: firstly, that this is the established usage and, secondly, that doing otherwise would risk circularity in claims about the process (Ritchie, 2006, p. 243). The definition of computational creativity offered by Wiggins (2006) referred to behaviour of systems which would be 'deemed creative if exhibited by humans' (p. 210). As late as 2012, Jordanous' definition referred to behaviour 'if observed in humans' (computationalcreativity.net, cited in Jordanous, 2012 p. 248). Although the Colton & Wiggins (2012, p. 21) definition quoted above in Section 2 does not make any reference to 'human', it does appear that many (if not all) other definitions made this comparison in some form. This distinction warrants further discussion from the computational creativity community.

Wiggins, Müllensiefen & Pearce (2010, p. 234) offer an interesting take on music and what it means in which they state: 'Music, in its own right, does not exist.' This refers to the fact that when we talk about 'music' what we are actually referring to is a specific representation of music such as an audio recording, a live show or a musical score. The only way in which these representations actually mean music to us is in our brains' interpretation of them. By this reasoning, ultimately music only exists in our minds. Although this may be a philosophical stance, it is an important one to consider from the outset when trying to establish how computationally generated music should be judged or evaluated. Taking this standpoint is one of the strongest arguments for continuing to use human judgements on generated music, if evaluation is purely performed on the artefact produced by the system.

On the other hand, discussions on the Creative Tripod in Colton (2008, p. 17) state that creativity can be exhibited by either the programmer, the computer or the consumer, thus asserting that creativity can be present regardless of the explicit opinion of the observer. Furthermore, Colton and Wiggins definition of computational creativity discussed in Section 2 foregoes any reference to human or human-like behaviour (Colton & Wiggins 2012, p. 21). As the field of computational creativity develops there appears to be a move away from describing or defining creativity in terms of human opinion. Should evaluation of systems developed in this field not follow such a move?

7.3 Who is the Music ‘For’?

When arguing against the use of human-based evaluation, the most obvious (and most often asked) question is ‘who is this music for?’ If we are suggesting that human evaluation is not the most important judgement (or, in the more extreme, not even relevant) to be made on autonomous music generation systems, then what is the point? Suggesting that such music is written for the enjoyment of computers is (certainly for the moment) silly, farcical and more suitable for weak science fiction than academic research. However, writing music for something else is not the goal or point of this research, nor is it something we currently aspire to. Fixating on ‘who’ the music is written for is again a pure judgement of the final artefact produced, rather than on the behaviour of the agent that created this music. Furthermore, it assumes that any future use of music must be interpreted by the same value-system as we have, an assumption that we may want to relax for a broader philosophical standpoint. We would suggest that the focus of this research is not to ask who the music is for but to completely disregard this notion of ‘for’ in an attempt to approach a more general evaluation of creativity involving a truly unbiased observer. Asking who the music is for is a natural question when considering music as a subjective, aesthetic and meaningful form of entertainment, but in this purely academic sense of considering computational creativity, a continued focus on human opinion is a meaningless distraction from the goal of unbiased evaluation.

8 Conclusion

This paper has presented a discussion on limitations that may arise when evaluating musical computationally creative systems. Evaluating creative systems inherently raises difficulties in that there is a subjective nature to the value of the artefact produced. In the case of musical systems, evaluating the output amounts to making an objective decision as to how ‘good’ the resultant piece of music is. This not only relies on a human definition as to what constitutes good music, but such tests only evaluate the final artefact produced by the system and not the behaviour of the system itself. Throughout this article we have discussed the implications of this and outlined possible limitations of considering generative music and creativity purely from this narrow standpoint. Section 2 introduced various types of creativity, computational creativity and some ideas that the field is built upon. Evolutionary methods applied to algorithmic composition were discussed in Section 3, including the difference between internally based fitness measures and external evaluation of the systems. Section 4 introduced the notion of conceptual space and the step-wise nature of creativity. The relationship between musicality, creativity and intelligence was discussed in Section 5. Previous methods of evaluating such systems were reviewed in Section 6. An overview of the implications of this research was discussed in Section 7.

We acknowledge that this discussion remains open-ended; we argue against limiting to human evaluation on music and creativity, yet recognise that this is still the logical way to evaluate such subjective systems. What we propose is that at this stage we open the discussion to the possibility that there may be alternative options, that aspiring to what we as humans think is best may not be the most general or most informative solution. We are currently in an age where AI is developing at a remarkable rate. If we are considering the capabilities of such AI systems in creative domains, we must surely broaden the possibilities within which to evaluate such capabilities. When we restrict evaluations to human-based judgement we may be assuming too much about systems whose limits and capabilities we are only yet discovering and which are growing and developing constantly. This is not a good time to limit the possibilities of any computational system.

The arguments presented here are not limited to the field of computer science. Never before have the boundaries between technology, art and philosophy been so vague or fluid. Pragmatically, it appears that making subjective judgements in comparison to what we know and believe as humans appears to be the only sensible option. Philosophically, however, we need to look to a broader picture. If Boden's vindication of Creative AI is to be realised, and if the Lovelace questions are to the answered, the argument for a more generalised evaluation of creative systems must be continued, regardless of whether they make us uncomfortable. After all, if the argument makes one uncomfortable or leaves one thinking of unanswerable questions, is that not what art, philosophy and technological development are all for?

References

Ariza, C. (2009). The interrogator as critic: The Turing test and the evaluation of generative music systems. *Computer Music Journal*, *33*(2), 48–70.

Biles, J.A. (1994). GenJam: A genetic algorithm for generating jazz solos. In *Proceedings of the international computer music conference* (pp. 131–137). Aarhus, Denmark: Michigan Publishing.

Biles, J. A. (2013). Straight-ahead jazz with GenJam: A quick demonstration. In Eigenfeldt, A., Bowne, O & Pasquer, P. (Eds), *Proceedings of the 2nd international workshop on musical metacreation* (pp. 20-23). Boston: Northeastern University.

Binsted, K., Pain, H. & Ritchie, G. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition*, *5*(2), 305–354.

Boden, M.A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*(1), 347–356.

Boden, M.A. (2004). *The creative mind: Myths and mechanisms*. Routledge, London and New York.

Boden, M.A. (2009). Computer models of creativity. *AI Magazine*, *30*(3), 23.

Brabazon, A., O'Neill, M., & McGarraghy, S. (2015). Grammatical evolution. In Brabazon, A., O'Neill, M., & McGarraghy, S. (Eds.), *Natural computing algorithms* (pp. 357–373). Berlin Heidelberg: Springer.

Bringsjord, S., Bello, P. & Ferrucci, D. (2001). Creativity, the Turing test, and the (better) Lovelace test. *Minds and Machines*, *11*(1), 3-27.

Burton, A. R., & Vladimirova, T. (1999). Generation of musical sequences with genetic techniques. *Computer Music Journal*, *23*(4), 59–73.

Campbell, M., Hoane, A.J. & Hsu, F. (2002). Deep blue. *Artificial intelligence*, *134*(1), 57–83.

Cardoso, A., Veale, T., & Wiggins, G. A. (2009). Converging on the divergent: The history (and future) of the international joint workshops in computational creativity. *AI Magazine*, *30*(3), 15.

Charnley, J., Pease, A., & Colton, S. (2012). On the notion of framing in computational creativity. In Maher, M.L., Hammond, K., Pease, A., Pérez y Pérez, R., Ventura, D. and Wiggins, G. (Eds.), *Proceedings of the third international conference on computational creativity* (pp. 77–81). Dublin, Ireland: University College Dublin.

Colton, S. (2008). Creativity versus the perception of creativity in computational systems. In Ventura, D., Maher, M.L. and Colton, S. (Eds.) *Proceedings of the AAAI spring symposium: Creative intelligent systems* (pp. 14–20). Palo Alta, California: The AAAI Press.

Colton, S. (2012). The painting fool: Stories from building an automated painter. In McCormack, J., d'Inverno, M. (Eds), *Computers and creativity* (pp. 3–38). Berlin Heidelberg: Springer-Verlag.

Colton, S., Pease, A. & Charnley, J. (2011). Computational creativity theory: The face and idea descriptive models. In Ventura, D., Gervás, P., Harrell, D.F., Maher, M.L., Pease, A. and Wiggins, G. (Eds.), *Proceedings of the second international conference on computational creativity* (pp. 90–95). Mexico City: Universidad Autónoma Metropolitana.

Colton, S. & Wiggins, G. A. (2012). Computational creativity: the final frontier? In De Raedt, L., Bessiere, C., Dubois, D., Doherty, P., Frasconi, P., Heintz, F. and Lucas, P. (Eds.), *Proceedings of the 20th European conference on artificial intelligence* (pp. 21–26). Montpellier, France: IOS Press.

Cook, M. & Colton, S. (2015). Generating code for expressing simple preferences: Moving on from hardcoding and randomness. In Toivonen, H. Colton, S., Cook, M., and Ventura, D. (Eds.), *Proceedings of the sixth international conference on computational creativity* (pp. 8-16). Utah: Bringham Young University

Cope, D. (2004). *Computer models of musical creativity*. Cambridge, Massachusetts and London: MIT press.

Dahlstedt, P. (2007). Autonomous evolution of complete piano pieces and performances. In Miranda, E.r, Martins, J. and Zhang, Q. (Eds.), *Proceedings of musicAL workshop*. Lisbon, Portugal: Springer.

Dahlstedt, P. (2009). Thoughts on creative evolution: a meta-generative approach to composition. *Contemporary Music Review*, *28*(1), 43–55.

de Freitas, A.R., Guimaraes, F.G. & Barbosa, R. V. (2012). Ideas in automatic evaluation methods for melodies in algorithmic composition. In *Proceedings of the sound and music computing conference* (pp. 514-520). Copenhagen, Denmark: Aalborg University Copenhangen.

Eigenfeldt, A., Bown, O., Brown, A. R. & Gifford, T. (2016). Flexible generation of musical form: Beyond mere generation. In Pachet, F., Cardoso, A., Corruble, V., and Ghedini, F. (Eds.), *Proceedings of the seventh international conference on computational creativity* (pp. 264-271). Paris, France: Sony CSL.

Eigenfeldt, A. & Pasquier, P. (2012). Populations of populations: composing with multiple evolutionary algorithms. In Machado, P., Romero, J., and Carballal, A. (Eds.), *Proceedings of the 1st International conference on evolutionary and biologically inspired music, art, sound and design* (pp. 72–83). Berlin Heidelberg: Springer-Verlag.

Fernandez, J.D. & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, 48, 513–582.

Goldberg, D.E., & Holland, J.H. (1988). Genetic algorithms and machine learning. *Machine Learning*, *3*(2), 95–99.

Hickinbotham, S., & Stepney, S. (2016). Augmenting live coding with evolved patterns. In Johnson, C., Ciesielski, V., Correia, J. and Machado, P. (Eds.), *Proceedings of the 5th International conference on evolutionary and biologically inspired music, art, sound and design* (pp. 31–46). Berlin Heidelberg: Springer-Verlag

Jordanous, A. (2011). Evaluating evaluation: Assessing progress in computational creativity research. In Ventura, D., Gervás, P., Harrell, D.F., Maher, M.L., Pease, A. and Wiggins, G. (Eds.), *Proceedings of the second international conference on computational creativity* (pp. 102-107) . Mexico City, Mexico: Universidad Autónoma Metropolitana.

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

Jordanous, A. (2013). Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application (Unpublished doctoral dissertation). University of Sussex.

JCMS (2017), *Journal of Creative Music Systems*. *Retrieved at* http://jcms.org.uk/docs/guidelines.html.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection* (Vol. 1). Cambridge and London: MIT Press.

Lehman, J., & Stanley, K. O. (2010). Efficiently evolving programs through the search for novelty. In Pelican, M. and Branke, J. (Eds.), *Proceedings of the 12th annual conference on genetic and evolutionary computation* (pp. 837–844). New York: ACM.

Loughran, R. & O'Neill, M. (2017). Clustering Agents for the Evolution of Autonomous Musical Fitness. In Correia, J., Ciesielski, V. and Liapis, A. (Eds.), *Proceedings of the 6th International conference on evolutionary and biologically inspired music, art, sound and design* (pp. 160-175). Berlin Heidelberg: Springer-Verlag

Loughran, R., McDermott, J. & O'Neill, M. (2015a). Grammatical evolution with Zipf's law based fitness for melodic composition. In Timoney, J. and Lysagh, T. (Eds), *Proceedings of the sound and music computing conference* (pp. 273-280). Maynooth, Ireland: Music Technology Research Group, Maynooth University.

Loughran, R., McDermott, J. & O'Neill, M. (2015b). Tonality driven piano compositions with grammatical evolution. In Obayashi, S., Poloni, C. and Murata, T. (Eds.), *Proceedings of the IEEE congress on evolutionary computation* (pp. 2168–2175). Sendai, Japan: IEEE.

Loughran, R., McDermott, J. & O'Neill, M. (2016). Grammatical music composition with dissimilarity driven hill climbing. In Johnson, C., Correia, J. and Ciesielski, V. (Eds.), *Proceedings of the 7th International conference on evolutionary and biologically inspired music, art, sound and design* (pp. 110-125). Berlin Heidelberg: Springer-Verlag

Loughran, R., & O'Neill, M. (2016a). Generative music evaluation: Why do we limit to 'human'? In Velardo V. (Ed.), *Proceedings of the 1st conference on computer simulation of musical creativity*. Huddersfield, UK: University of Huddersfield.

Loughran, R., & O'Neill, M. (2016b). The popular critic: Evolving melodies with popularity driven fitness. In Pasquer, P., Bowne, O & Eigenfeldt, A. (Eds), *Proceedings of the 4th international workshop on musical metacreation*. Paris.

Manurung, R., Ritchie, G., Pain, H., Waller, A., O'Mara, D. & Black, R. (2008). The construction of a pun generator for language skills development. *Applied Artificial Intelligence*, *22*(9), 841–869.

Marsden, A. (2000). Music, intelligence and artificiality. In Miranda, E. (Ed.), *Readings in Music and Artificial Intelligence* (pp. 15– 28). Amsterdam: Harwood.

McCormack, J. (2005). Open problems in evolutionary music and art. In Rotlauf, F. Branke, F., Cagnoni, J., Corne, S., Drechsler, D.W., Jin, R. Machado, Y., Marchiori, P. Romero, J., Smith, G.D. and Squillero, G. (Eds.), *Proceedings of workshops in applications of evolutionary computing* (pp. 428–436). Berlin Heidelberg: Springer-Verlag.

McFarland, D. (2009). *Guilty robots, happy dogs: the question of alien minds*. Oxford: Oxford University Press.

Miranda, E.R. (2004). At the crossroads of evolutionary computation and music: Self-programming synthesizers, swarm orchestras and the origins of melody. *Evolutionary Computation*, *12*(2), 137–158.

Moroni, A., Manzolli, J., Von Zuben, F. & Gudwin, R. (2000). Vox populi: An interactive evolutionary system for algorithmic music composition. *Leonardo Music Journal*, *10*, 49–54.

Munoz, E., Cadenas, J., Ong, Y. S. & Acampora, G. (2016). Memetic music composition. *IEEE Transactions on Evolutionary Computation*, *20*(1), 1–15.

MuMe (2017) musical metacreation. Retrieved from http://musicalmetacreation.org

Pearce, M., Meredith, D. & Wiggins, G. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientiae*, *6*(2), 119–147.

Pearce, M. & Wiggins, G. (2001). Towards a framework for the evaluation of machine compositions. In Colton, S. (Ed.), *Proceedings of the AISB01 symposium on artificial intelligence and creativity in the arts and sciences* (pp. 22–32). York: University of York

Pearce, M. T., & Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In Cardosa, A. and Wiggins, G. (Eds.), *Proceedings of the 4th international joint workshop on computational creativity* (pp. 73–80). London: Goldsmiths University

Reddin, J., McDermott, J., & O'Neill, M. (2009). Elevated Pitch: Automated grammatical evolution of short compositions. In Giacobini, M. Brabazon, A., Cagnoni, S., Ekart, A., Esparcia-Alcázar, A.I., Farooq, M., Fink, A., Machado, P., McCormack, J., O.Neill, M., Neri, F., Preuss, M., Rothlauf, F., Tarantino, E., Yang, S. (Eds.), *Proceedings of workshops in applications of evolutionary computing* (pp. 579–584). Berlin Heidelberg: Springer-Verlag.

Ritchie, G. (2001). Assessing creativity. In Colton, S. (Eds.) *Proceedings of the AISB01 symposium on artificial intelligence and creativity in the arts and sciences*. York: University of York

Ritchie, G. (2006). The transformational creativity hypothesis. *New Generation Computing*, *24*(3), 241–266.

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

Scirea, M., Togelius, J., Eklund, P. & Risi, S. (2016). Metacompose: A compositional evolutionary music composer. In Johnson, C., Correia, J. and Ciesielski, V. (Eds.), *Proceedings of the 7th International conference on evolutionary and biologically inspired music, art, sound and design* (pp. 202–217). Berlin Heidelberg: Springer-Verlag.

Shao, J., McDermott, J., O'Neill, M. & Brabazon, A. (2010). Jive: A generative, interactive, virtual, evolutionary music system. In *Proceedings of workshops in applications of evolutionary computing* (pp. 341–350). Berlin Heidelberg: Springer-Verlag.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G. *et al*. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Spencer, R. (2012). The size and shape of "idea space". *International Journal of Innovation Science*, *4*(2), 71–76.

Stanley, K.O., & Lehman, J. (2015). *Why greatness cannot be planned: The myth of the objective*. Switzerland: Springer.

Todd, P.M., & Werner, G.M. (1999). Frankensteinian methods for evolutionary music. In Griffith, N. and Todd, P.M. (Eds.), *Musical networks: parallel distributed perception and performance* (pp. 313-340). Cambridge and London: The MIT Press.

Valitutti, A., & Veale, T. (2016). Infusing humor in unexpected events. In Streitz, N. & Markopoulos, P. (Eds.), *Proceedings of the international conference on distributed, ambient, and pervasive interactions* (pp. 370-379). Switzerland: Springer International Publishing.

Waschka II, R. (2007). Composing with genetic algorithms: GenDash. In Miranda, E.R. and Biles, J.A. (Eds.), *Evolutionary computer music* (pp. 117–136). London: Springer-Verlag.

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

Wiggins, G.A., Müllensiefen, D. & Pearce, M.T. (2010). On the non-existence of music: Why music theory is a figment of the imagination. *Musicae Scientiae*, *14*(1_ suppl), 231–255.