What Makes Machine Music Machine Like?

Group Member Names: Benjie Genchel, Jared Henson, Yuqi Cao

Introduction

Can humans truly tell the difference between music composed by humans and music generated by a modern deep learning based statistical model? Without the bias of knowing whether the music they hear is composed by a machine or a human, what characteristics would people identify as human like, or machine like?

Symbolic music generation using deep learning based models has become increasingly popular over the past several years, and interest in the field continues to grow. However, despite a proliferation of creative approaches and claims of improved performance, no model has been able to produce substantially human-like music. That is to say, generally speaking, it appears intuitive that it would be easy for most people to differentiate a piece of music produced by a human from one produced by a deep learning model. One need only listen to a substantially long generation from any of these systems to intuit that the layman who inevitably asks if researchers are attempting to replace human musicians has nothing to fear at present, regardless of the honest response. In the event that making such a classification is in fact as easy as it appears, we would not know for certain, as the majority of evaluations presented in this area ecompare their model's output to a set of baseline models along a set of objective (though contestable) metrics or via a listening test in which each contestant is rated along preset dimensions. Thus, there has been little work in the way of comparing machine generated music to actual music; characterizing the perceptual differences between the two and discovering how people listen and what they listen for in the context of these studies. In this study, we seek to test whether humans can tell the difference between music generated by a modern state of the art deep learning system and music composed by a human being, as well as gather qualitative data on how such a distinction is made; what characterizes composition by a human vs. composition by a machine?

We hypothesize that, within the parameters under which modern music generation models have been deemed successful (at a short time scale), humans would not be able to tell a difference between music generated by a model and music composed by a human. We will compare between a model trained on jazz music, generating over chords from Jazz standards and the original melodies that fit with those chords. We hope to obtain more sanitized and interesting results in this manner, as we believe Jazz standards are less familiar to most than pop or classical pieces, and contain complex rhythmic, melodic and harmonic patterns which prevent a simple repetition of notes in key from being mistaken as human.

As such, we additionally hypothesize that we will find significant differences in the identifications and discriminatory ability of subjects who are experienced with our chosen genre.

Background

Music generation has long been a target task of machine learning, dating back to the first recurrent neural networks developed during the 1990s. Its popularity has only grown, and is currently at a peak in the wake of the resurgence of deep learning and an increasing desire for easily obtainable, custom thematic and background music for various multimedia projects [2]. Long-Short Term Memory recurrent neural networks (LSTM-RNNs), Convolutional neural networks (CNNs) and Generative Adversarial Networks (GANs) are favorites for sequential generation, while increasing attention has been focused towards embedding space models such as Variational Autoencoders (VAE). Colombo et al. train two separate networks for pitch and duration of notes for monophonic melody generation, conditioning the generation of durations on the next occurring pitch in the training sequence [6]. In Songs from PI, the authors attempt to generate multi-track pop music via a series of hierarchically conditioned LSTMs, with the lowest level generating a drum track and the highest producing harmony [4]. Yang et al.'s MIDINet presents a combination of CNNs and GANs for the purpose of generating pop music in piano roll format [5]. Google Magenta's MusicVAE presents a recent attempt at generating longer form music using neural networks via conditioning of a group of RNNs in sequence using an embedding of long form music [3]. Vox Populi takes a different approach altogether, though one still under the umbrella of machine learning, employing an evolutionary algorithm to evolve a population of chords based on a fitness function that can be interactively manipulated by a user [1].

While many of these efforts are evaluated by listening tests, the value of the ways in which these tests have been carried out have been inconsistent across research efforts, and frequently critiqued. Ariza discusses the use of the term turing test in the context of evaluating generative music systems, arguing that the majority of uses of the term 'Turing Test' to describe evaluations in this domain are false; the tests that were conducted do not fit into the Turing framework [7]. Pease and Colton argue that it is currently untenable to apply a Turing-style test because the Turing-style tests that can be feasibly carried out in this domain lack meaningful participant interaction and do not take creative motivation and direction into account [8]. Even further, there has been little work in the way of exploring qualitatively how this music is being judged. Previous research in this field has indicated that humans are largely able to tell the difference between automated or generated compositions versus those of other people. Using jazz as the testing material may provide interesting results, as the improvisation of jazz players is less predictable. Chella and Manzotti support this notion, proposing a form of Turing test using a computer agent that produces jazz improvisations. In addition, they laid out their requirements for such a system to perform at the level of an expert jazz improviser [14].

In Colton's study, they review some existing frameworks for assessing artifact generation programs [9]. They conclude that an ideal software considered to be creative needs to subvert any given notions of good and bad artifacts. Jordanous proposed a standardisation procedure for evaluating creative systems [10], consisting of three steps: stating what it means for a computational system to be creative, deriving and performing tests based on these statements. Pearce and Wiggins presented a method for evaluating creative systems based on Markov model learning [11]. The evaluation model is learning-based and can be enhanced by applying mathematical methods over data, from which our research could learn.

Chamberlain, Mullin, and Wagemans conducted a similar experiment to our proposed one with generated visual images as opposed to auditory stimulus. They found that individuals were able to discern the machine images from the human ones. Interestingly, they also found that works categorized as man-made had a much higher aesthetic rating from the participants [12]. The Chamberlain group also revealed, in a different study, an explicit prejudice against computer generated art in a testing environment. This prejudice was partially reversed, however, if the viewers saw the robot and could infer anthropomorphic characteristics from it [15].
This may indicate some bias introduced by asking for the classification of the piece before asking for an aesthetic rating. Boden at the University of Sussex found that the Turing test for computer art has been passed behaviorally before, also examining the link between computer art and the lack of computer creativity [13].

Methodology

In order to assess our subjects' ability to differentiate human composed melodies from machine generated melodies, we conducted a within-subjects design experiment, in which we grouped participants based on familiarity with Jazz. We had initially planned to additionally group participants by musical proficiency, but nearly all our participants were highly proficient. Our experimental procedure consists of an online survey in which participants were asked to listen to and then make verbal and ordinal observations on the perceived humanity or inhumanity of pairs of model generated and human composed melodies played over over a shared chord progression.

First, each subject was asked to state their musical background/proficiency and familiarity with Jazz music, Then, they were asked to sequentially complete four trials of a listening and appraisal task. During each trial, subjects were exposed to a pair of short audio segments, each consisting of a monophonic melody played over chords and a baseline. Both segments shared the same chords and duration, however one contained a melody generated by a deep learning model, and the other the original melody corresponding to those chords.

Audio segments were each rendered identically from symbolic MIDI, with the melody played by a high quality saxophone sampler, the chords played by a high quality piano sampler, and the baseline (derived from the chords via arpeggiation) played by a high quality stand up bass sampler. Segments in each trial ranged from 5 to 10 seconds, and were presented in the

form of youtube videos, which were embedded in the form and contained no information save for their corresponding trial and melody number.

After listening to each segment of audio, subjects were presented with questions regarding how they perceived what they heard. Specifically, for each segment, subjects were asked to rate the perceived human-ness and machine-ness on a scale from 1 to 5, as well as describe qualitatively what they felt were human or machine-like properties. At the end of each trial, the subjects were asked to identify which one of the two melodies they believed was composed by a human, how confident they were in that decision, and to describe how they came to their conclusions.

The artificially composed segments were generated by a Long Short Term Memory Recurrent Neural Network (LSTM-RNN) based deep learning model, consisting of two individual networks separately modeling pitch and duration of notes. During training, each is conditioned with the other's data, along with corresponding chords and position within the current bar. All of these are time aligned, and fed into the models one note at a time. During inference, the two run concurrently, feeding their outputs into the other's input for conditioning, with conditioning chord transitions and bar positions calculated on the fly according to generation length. Each individual network was trained for 100 epochs with a learning rate of 1e-3 using the Adam optimizer on a collection of around 200 Bebop Jazz lead sheets in musicxml format. A diagram of the model can be seen below (Fig. 1).

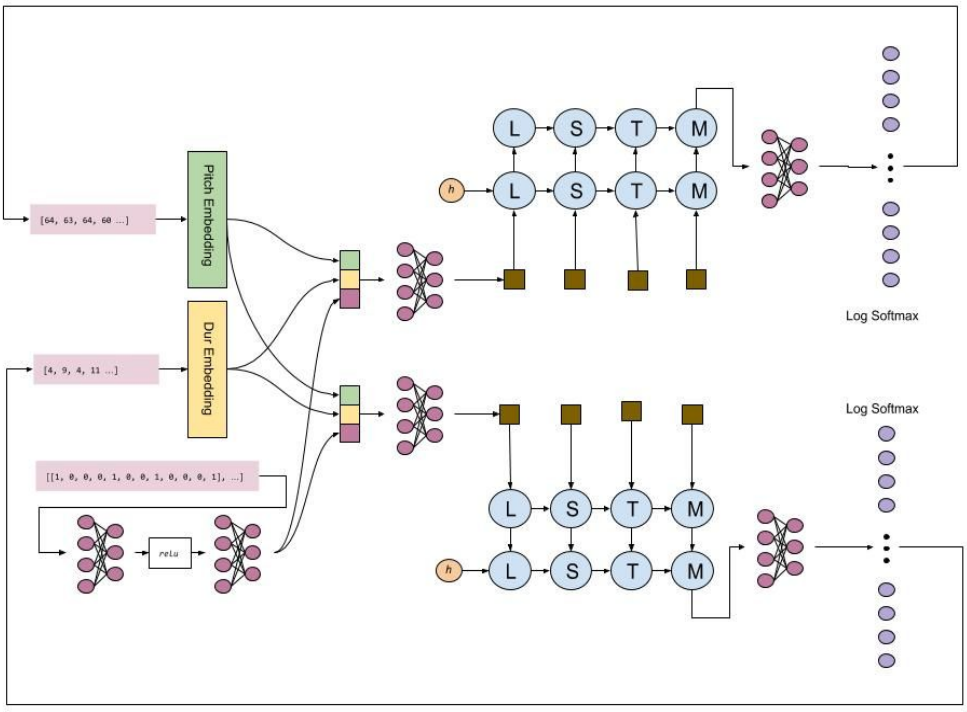

Fig. 1: Diagram of Neural Network Model used for Music Generation

Results

We first tested the general hypothesis that, at the scale at which we presented our stimuli, subjects would not be able to perceive a difference between generated and composed segments. We used a binomial test to evaluate across all trials and subjects, with hypothesis p = 0.5. The p_value calculated from the test was near **1e-6**, showing that our assumption of random guess level performance was heavily inaccurate. The subjects gave 46 correct judgements out of 56 total trials.

Next, we used an independent t-test to evaluate the difference between the discerning ability of subjects who were familiar with jazz and those who were not. We felt an independent t-test was appropriate here because the representing values were either correct or incorrect as represented by a 1 or 0 respectively, allowing the mean of all observations within a group to reflect the overall accuracy of that group on a continuous scale. Additionally, since all subjects took the test independently, we can assume that the variance across each population would be equal. The result of the test was a p_value of **0.367**, far above the typical threshold of 0.05. This shows that there was no statistically significant difference between each group's ability to discern between human composed and machine generated music.

The qualitative descriptions given by the participants were wide ranging in the qualities they focused on and the specificity with which they could articulate what they were perceiving. In order to simplify the qualitative data and make it available for analysis, we went through it by hand and assigned each description a set of relevant tags which captured the essence of what the participants were addressing (Table 1). It should be noted that tags defined the topic of focus, not the valence. For example, 'Naturality' was not only given to writings that indicated the naturality of a segment, but also to writings which indicated unnaturality.

Tag	**Description**
Harmonicity	How well notes in the melody fit with the backing track; how well the melody fit with the chords.
Improvisatory	How much like improvisation the segment sounded.
Interestingness	How interesting, challenging, unique, or 'fun' a melody sounded.
Naturality	A blanket term for indescribable, non specific feelings that subjects perceived in the music. Emotion, humanness, expressivity, and general pleasantness all fall under this category.
Nothing	This tag describes the frequent occurrence of subjects writing that they felt either nothing was human about a piece or nothing was artificial about a piece. In its ambiguity, it could be considered part of 'Naturality'.
Pitches	Dealing with the pitch contour, jumps in pitch between notes, general pitch choices, etc. Anything to do with the pitch choices of the melody, independent of the chords.

Recognize	This tag indicates that the subject recognized the music they were hearing, and so evaluated based on that recognition instead of blind perception.
Rendering	This tag indicates that the subjects based their judgements and appraisals based on the audio itself, including the articulation of the instruments, the perception of how the 'player' was using the instrument, the changes in note velocity, etc.
Rhythm	Dealing with note spacing, rhythmic pattern, syncopation, meter and other local features independent of longer term or overall structure.
Structure	Dealing with phrasing, perceived overall direction and intent, inference into how the segment would continue after its end, repetition, and longer term structure.

Table 1: The set of tags used to represent what subjects focused on when deciding and appraising stimuli

The total distribution over the tags, as well as the distributions for subjects familiar with jazz and those unfamiliar are displayed below (Fig. 2)

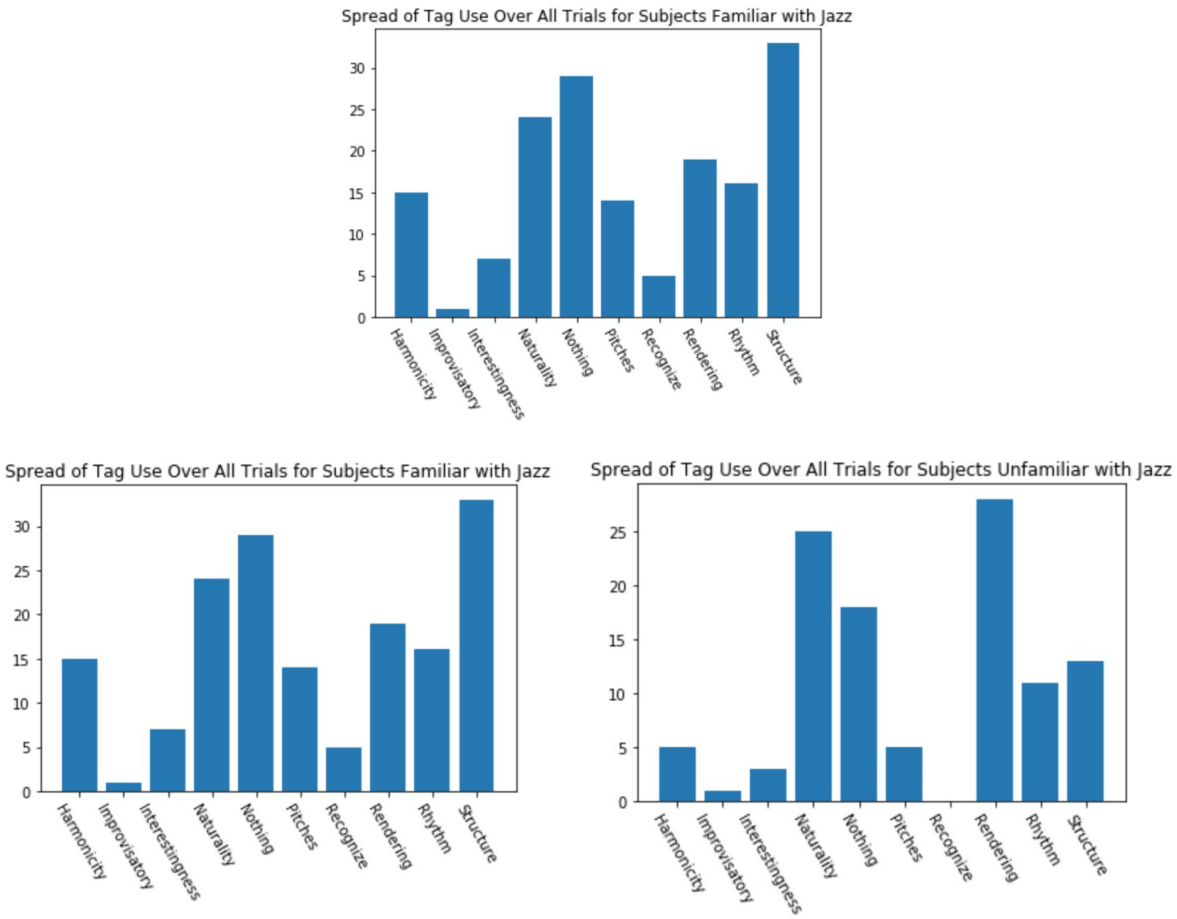

Fig 2. Distributions of description tags over all trials for all participants, the familiar with jazz group and the unfamiliar with jazz group.

Finally, we compared the human-likelihood and machine-likelihood ratings given by subjects between the two familiarity groups on four categories which are implicit in our methodology: likelihood human composed is human composed, likelihood human composed is machine generated, likelihood machine generated is human composed, and likelihood machine generated is machine generated. We compared each group on their distribution on each category using an independent t-test, which also works in this situation because the ratings are ordinal and the populations can once again be assumed to be equal in variance. The results of the t-tests are displayed in Table 2.

Category	p_value
Likelihood Human Composed is Human Composed	0.0002089169619897932
Likelihood Human Composed is Machine Generated	0.7139506729852861
Likelihood Machine Generated is Human Composed	0.22402856405063756
Likelihood Machine Generated is Machine Generated	0.2986319848078392

Table 2: Independent t-test results comparing the jazz familiar group with the jazz unfamiliar group in terms of how human-like or machine-like they rated each segment.

The results in Table 2 show that the null hypothesis holds for each category, meaning no statistically significant difference in the ratings between groups, except the first - rating how likely it is that actual human composed music segments were composed by a human. The ratings distribution for each group for this category are displayed in Fig. 3.

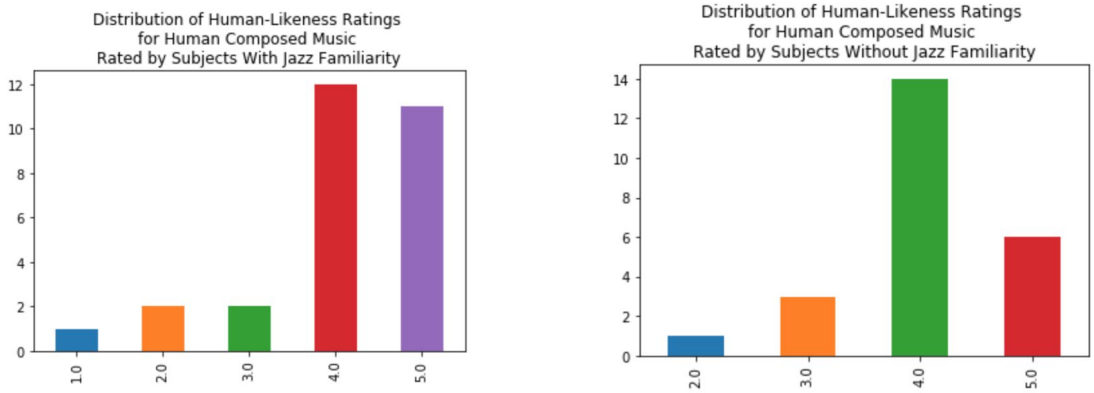

Fig. 3: Distributions of ratings describing how likely subjects believed the segment they heard was composed by a human being for the case that the segment they had heard/were judging was actually composed by a human. An independent t-test shows a statistically significant difference in rating between the populations.

Conclusion

In this experiment, the derived results indicated both our original hypotheses to be incorrect: the ability of our subjects to discern human composed from machine generated music was shown to be far greater than expected, as shown by binomial test, the distribution of correct and incorrect answers was far divorced from a uniform distribution. The ability of subjects

familiar with jazz as compared to those who were not to discern human composed and machine generated music was additionally shown to be false by an independent t-test.

Distributions over assigned tags between groups were shown to be similar, indicating even propensities towards perception and differentiation. Independent t-tests between groups with respect to ratings of segments as human-like and machine-like showed highly correlated results except in the category of rating human composed music as human composed. In this category, the t-test showed a significant difference. Plots of the ratings distributions for the two populations indicate that this could be due to a greater degree of certainty among the jazz familiar group, which is sensible.

Small sample size and poor sample representation (all samples were students in a music program) prevented us from analyzing the data with respect to general musical ability (all participants were musically proficient) or confirming the generality of our results here. However, the data we were able to gather through our experiments showed a great degree of depth and is ripe for further exploration.

The question of how to evaluate creative systems is a continual problem in all computational art, and certainly in the area of symbolic music generation with deep learning systems. While we once again acknowledge our lack of ability to make strong claims given our sample size and representation, by eschewing the traditional comparative evaluation schemes common in the field and allowing subjects to provide detailed qualitative data alongside ordinal ratings, we were able to inch a closer towards a clearer understanding of the differences between current generated music and the goal of generative systems - human like music, and how those differences are perceived.

Discussion

The point of this experiment was to see how subjects would perceive machine generated music without the context or knowledge that it was in fact machine composed. We had hoped to structure the experiment such that we would avoid the good-participant and bad-participant demand characteristics and maintain the impartiality of the subject in each trial. However, it was clear from our results that the way the survey was structured immediately allowed for just such a bias. Because the questions in each trial were the same, and both stimuli were shown to the subject at the same time, it appears that it was easy to begin immediately making a mental decision on which was which, because it was known that one was machine generated and the other composed, before even beginning to answer the questions. We believe this is a possible explanation for why so many ‘nothing’ tags (nothing about this was human-like, nothing about this was artificial) appeared in the experiment, even when the machine generated music was specifically chosen due to its consonance and human-like affect.

Additionally, it appears that the description of the task was not well understood by some participants. The point was to judge the melodies, with the sampler rendering only used to even the playing field and give human-like feeling to both stimuli in the pair. Yet, so many participants made judgements based on the sound quality, dynamics, and other effects which had nothing to do with the model. Further, there were a few subjects whose final decisions did not appear to match how they rated human-likeness and machine-likeness earlier in the trial.

We believe that the qualitative descriptions were an important part of the experiment, and one which could with more data and time, contribute many interesting finds, however, it was difficult to interpret the responses. Many were ambiguous, and did not address specific features of the music, only what the subject seemed to feel or intuit. It is tempting to, like many other researchers, restrict the input from subjects via predefined categories and ratings, however, this would hinder the possibility of receiving novel observations and insights. Surely, it is quite difficult for most people to describe with specificity features of something abstract like music; we will certainly consider how to guide these qualitative responses and better convey what we are looking for if this experiment is run in the future.

We feel that this experiment was a success overall. The idea of collecting this type of data feels simple and straightforward when standing distant at the point of conception; getting our hands dirty and actually seeing how people respond in this type of situation really brings into perspective the challenges one faces in developing and evaluating generative systems. It is easy to intuit how people act based on how one thinks, however, it is enlightening to see how different people can be, and how differently people can think about the same things. Though we didn't receive the data we were expecting, we did receive a lot of interesting information about experimental design and how people perceive the data, which was a larger, more general goal.

Future Work

To further the work in this report, it would be most useful to examine which specific musical qualities affect the listener's perception of a melody as human or machine composed. Factors such as note length and quality, rest length, and overall tonicity may play a key role in informing participant decision. This experiment focused primarily on the relationship between the participants' previous familiarity with jazz music and their ability to distinguish the melodic clips, without considering elements of the music itself. If a study was designed that controlled for musical genre in some way, various musical qualities could be manipulated in both human and generated melodies in order to gauge effect on the participant. The musical key of each melody may also factor into user perception. If all musical clips were restricted to the same key, it may shed insight into how much of the participants' decisions were based on a sense of dissonance among the notes in a single melody.

In addition to participant identification of machine-composed melodies, aesthetic considerations are also frequently measured in Turing-test focused studies. Previously mentioned musical descriptors may also influence aesthetic perception of the melody, perhaps independent of its classification as human or machine composed. A study conducted separately from a Turing-test may provide more clear insight, especially if participants were unaware of the origins of the melodies.

References

1. Moroni, Artemis, et al. "Vox populi: An interactive evolutionary system for algorithmic music composition." *Leonardo Music Journal* (2000): 49-54. Accessed 12 Sept. 2018.
2. Jean-Pierre Briot, Gaetan Hadjeres, and Francois Pachet. Deep learning techniques for music generation - A survey. CoRR, abs/1709.01620, 2017. Accessed 12 Sept. 2018.
3. Roberts, Adam, Jesse Engel, and Douglas Eck. "Hierarchical variational autoencoders for music." *NIPS Workshop on Machine Learning for Creativity and Design*. 2017. Accessed 12 Sept. 2018
4. Hang Chu, Raquel Urtasun, and Sanja Fidler. “Song from PI: A musically plausible network for pop music generation.” *ICLR Workshop,* 2017. Accessed 20 Aug. 2018.
5. Yang, Li-Chia, Chou, Szu-Yu, Yang, Yi-Hsuan. “MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation.” *Conference of the International Society for Music Information Retrieval,* 2017. Accessed 20 Aug. 2018.
6. Colombo, Florian, Alexander Seeholzer, and Wulfram Gerstner. "Deep artificial composer: A creative neural network model for automated melody generation." *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, Cham, 2017. Accessed 12 Sept. 2018.
7. Ariza, Christopher. "The interrogator as critic: The turing test and the evaluation of generative music systems." *Computer Music Journal* 33.2 (2009): 48-70. Accessed 12 Sept. 2018.
8. Pease, Alison, and Simon Colton. "On impact and evaluation in computational creativity: A discussion of the Turing test and an alternative proposal." *Proceedings of the AISB symposium on AI and Philosophy*. 2011. Accessed 12 Sept. 2018.
9. Colton, Simon. "Creativity Versus the Perception of Creativity in Computational Systems." *AAAI spring symposium: creative intelligent systems*. Vol. 8. 2008. Accessed 12 Sept. 2018.
10. Jordanous, Anna. "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. Accessed 12 Sept. 2018
11. Pearce, Marcus T., and Geraint A. Wiggins. "Evaluating cognitive models of musical composition." Proceedings of the 4th international joint workshop on computational creativity. Goldsmiths, University of London, 2007.

12. Chamberlain, Rebecca, Caitlin Mullin, and Johan Wagemans. "The artistic Turing test: An exploration of perceptions of computer-generated and man-made art." *Journal of vision*15.12 (2015): 112-112.
13. Boden, Margaret. "The Turing test and artistic creativity", *Kybernetes*, Vol. 39 Issue: 3, (2010): 409-413. Accessed 29 September 2018
14. Chella, Antonio, and Riccardo Manzotti. "Jazz and machine consciousness: Towards a new Turing test." *Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World* (2012): 49. Accessed 29 September 2018
15. Chamberlain, Rebecca; Mullin, Caitlin; Scheerlinck, Bram and Wagemans, Johan. 2017. Putting the Art in Artificial: Aesthetic Responses to Computer-generated Art. *Psychology of Aesthetics, Creativity and the Arts*, ISSN 1931-3896 [Article]