UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di laurea triennale in Scienze Psicologiche Cognitive e Psicobiologiche

When Music Meets Machine:
Perceptual Analysis of AI Music.

***Relatore*:**
Prof. Massimo Grassi.

***Laureando*: Vito Piccione**
***Matricola:* 2013352**

Anno Accademico: 2022-2023

Contents

Introduction

In the last few years there has been an evident surge in the use of AIs among the general population. The public release of image generators powered by AI (DALL-E, MidJourney...) allowed the general public to produce impressive images starting from a simple text prompt. Social networks quickly flooded with AI images and, suddenly, a powerful spotlight was shining on AIs. In this newfound attention, highly proficient language models started to claim their place (ChatGPT, Elicit...), offering to aid us in research, writing, coding and so much more. In this rapidly evolving environment, there has been much talk about the value of AI-generated content.

This paper focuses on an aspect often overshadowed by the pleasantness of AI art: recognition. In what can be assimilated to an artistic Turing Test, this paper aims to determine whether listeners can distinguish between AI-generated music and music composed by humans. Building upon recent experiments that have revealed a positive correlation between high musical expertise and enhanced perceptual abilities in music (Castro and Lima, 2014), this paper also investigates whether specific characteristics such as music sophisticatedness and familiarity with a particular music genre can impact the ability to discriminate between AI-generated music and human-composed music. In an effort to break away from the tradition of classical music as a staple of music psychology, a more holistical approach was used, considering multiple genres and instruments.

1 Theorethical Background

1.1 The Rise and Fall of AIs: A Brief Lesson on AI History.

Mythology of Artificial Intelligence

While the concept of artificial intelligence (AI) as we understand it today emerged with the advent of computers, the notion of creating intelligent beings can be traced back to ancient mythologies. Mythologies across different cultures often featured tales of humans constructing or creating intelligent entities, even before the concept of electronic computing came into existence. These myths embody the human fascination with crafting beings that possess intellect and abilities beyond ordinary humans.

In Greek mythology, for instance, the god Hephaestus, known for his skill as a blacksmith, fashioned remarkable mechanical servants. One such creation was Talos, a giant bronze automaton tasked with protecting the shores of Crete.

Similarly, in Jewish folklore, the Kabbalistic tradition speaks of the golem, a creature molded from clay. The golem is often portrayed as an obedient servant, created through mystical rituals and inscribed with sacred words or symbols. These legends depict humans imbuing life into an artificial being, a construct that could perform tasks and fulfill specific purposes.

These mythological accounts provide glimpses into the long-standing human desire to bring artificial beings to life, predating the technological achievements of AI by centuries or even millennia.

Formal Reasoning

Before the birth and spread of formal reasoning, the precursors of AI had a general essence of imitating human behaviour. Both Talos and the Golems didn't involve thought in any way that mattered.

When the concept of formalized human thought arose, these deeply-rooted images started to take on a new form.

Ramon Llull (1232-1316), a philosopher and theologian from the thirteenth century, developed a system called the Ars Magna (Great Art) which aimed to mechanize human reasoning (Fidora et al., 2011). Llull's system involved the use of combinatorial logic and symbolic notation to generate and analyze arguments systematically. He believed that through the manipulation of symbols and logical operations, one could arrive at universal truths (Crossley, 2005).

A few centuries later Gottfried Wilhelm Leibniz (1646-1716) contributed significantly to both mathematics and philosophy, including the development of formal systems and logical reasoning. Leibniz envisioned the possibility of a universal language or calculus that would represent all human knowledge in a formal and logical way. He proposed the idea of a "calculus ratiocinator," a symbolic system that could mechanically manipulate concepts and reason deductively.

The contributions of Llull and Leibniz in formalizing human reasoning and envisioning mechanical systems for manipulating symbols and concepts were instrumental in shaping the concept of AI.

Computers

During the first half of the twentieth century, a remarkable convergence of theories and discoveries occurred. The discovery of the brain as an electrical network, made of neurons that fire in all-or-nothing signals, firstly by R. Caton (1842-1926) in 1875 and then throughout the following century by researchers such as H. Berger (1873-1941), K. Brodmann (1868-1918), C. Golgi (1843-1926), Santiago Ramón y Cajal (1852-1934); Shannon's description of digital signals as binary (Shannon and Weaver, 1949); Alan Turing's theory of computation, affirming that any form of computation could be described digitally (Turing, 1937). From these ideas, the concept of a completely electronic brain started to emerge, more approachable than ever.

In 1943 Walter Pitts and Warren McCulloch decided to address the situation, building the theory for what could be described as the first neural network: a theoretical system capable of performing simple logical functions by stringing together idealized artificial neurons (McCulloch and Pitts, 1943). This first attempt to describe a working electronic brain sent a powerful wave of inspiration through young students. Among them was a young Marvin Minsky, a 24 years old graduate that would continue on to build, only seven years later, the first neural net machine and become a leading figure in the AI field for the 50 years to come.

Another important development that took place in these years was the publishing of Alan Turing's paper that formalized the Turing Test (Turing, 1950). If a machine could carry on a conversation that was indistinguishable from one with a human being, it was within reason to say that the machine was thinking. Albeit this being a rather controversial conclusion, it provided researchers with a clear goal and an operationalization of the problem.

Despite these early steps, the field of AI wouldn't come to be until the Dartmouth Workshop in 1956 (Muthukrishnan et al., 2020). Organized by Minsky and other leading researchers, it was here that the term Artificial Intelligence was first proposed. As stated by J. Moor (2006) in its interview with five of the researchers that attended the workshop, there was no accord on a unifying theory; the field originated not from an agreement on methodology, but rather from the shared vision of computers performing intelligent tasks.

After the workshop took place, the field of AI started to collect successes. In rapid development, computers started to solve increasingly difficult logic problems (ie. General Problem Solver, built by Herbert A. Simon, J. C. Shaw, and Allen Newell in 1959) and learn to recognize and use english speech (Shoebox computer built by IBM in 1962, ELIZA chatbot by Joseph Weizenbaum in 1965). In a mistake that would prove a bad habit of the field, researchers expressed an intense optimism, predicting the creation of a fully intelligent machine in less than 20 years (Simon, 1965). The overwhelming optimism, coming from a context of rapid development, drew in large amounts of funds, both from privates and from governments.

Eventually the field started to slow down, weighted by many problems that had been underestimated, namely computer power and the commonsense knowledge problem (imparting AI systems with the intuitive understanding of everyday knowledge). When this happened fundings rapidly stopped (Crevier, 1993). The six years to follow (1974-1980) would become known as the first AI winter: hopes were low, fundings were even lower.

In 1980 however, through a reframing of the situation, the AI field started to regain some traction. Leaving the idea of a general intelligence system behind, researchers started to develop a form of AI program called “expert system”. This new paradigm allowed researchers to create a highly specialized program, capable of answering questions or solving problems about a specific domain of knowledge. This form of AI looked promising, especially for corporations around the world because of its possible use in automating high-knowledge tasks.

Edward Feigenbaum was the first to develop such a system, trying to automate the identification of compounds from spectrometer readings (Feigenbaum and Buchanan, 1993). Another expert system capable of diagnosing infectious blood diseases was developed a few years later.

The main difference between expert systems and the previous wave of AI programs was that the former were easier to build and, more importantly, had practical, useful applications (Crevier 1993, pp. 158–159). Furthermore, by focusing on a specific field, the commonsense knowledge problem could be avoided. As a matter of fact, the most relevant expert system (XCON) allowed its company to save 40 million dollars annually (Crevier 1993, p. 198).

The conceptual framing of expert systems also led to the birth of Cyc, the first attempt to tackle the commonsense knowledge problem. Led by Douglas Lenat, Cyc was supposed to be a massive database containing all the mundane facts the average person knows. The project was of course a vast undertaking and was not expected to be completed in a short time (McCorduck, 2004, p. 489).

In 1981 a massive investment of the Japanese ministry of international trade and industry set off a chain investment in the field of AI both nationally and internationally, rekindling the interest in the field (McCorduck, 2004, pp. 436-441).

In the span from 1981 to 1987, the business community’s infatuation with AI followed a familiar pattern of an economic bubble. The initial warning sign of the decline emerged when the market for specialized AI hardware experienced a sudden collapse in 1987. Moreover, maintaining successful expert systems like XCON proved to be excessively costly. As a result, by the conclusion of 1993, more than three hundred AI companies had ceased operations, either due to bankruptcy or acquisition (Newquist, 1994, p. 440).

The time frame that extends from 1987 to 1993 is now considered to be the second AI winter. However, even with low fundings, the field effectively reached many goals in this time. In 1989, for example, two different AIs won against chess masters. One of them, DeepThought, would become the prototype for DeepBlue. In 1990 the first search engine was developed and the chatbot Cleverbot was launched. Furthermore a new tendency came up in these years. The symbol processing model was being doubted by many researchers that argued for a more body-focused cognition, an holistic approach in which mind and body are impossible to separate (Brooks, 1980). This would become known as the Embodied Mind Thesis.

In 1993 the field of AI presented itself as deeply fragmented, with a bad reputation in the business world (Newquist 1994, pp. 511). In spite of the uninspiring setting, things were starting to change. Technological progress was giving the AI field systems powerful enough to tackle the intrinsic computation power problem. These technological advancements made many achievements possible. In 1997 Deep Blue became the first computer chess-playing system to beat a reigning world champion (McCorduck, 2004, pp. 480-483; Schaeffer and Plaat, 1997). It was estimated to evaluate 200 million positions per second.

In 2005 a Stanford robot won the DARPA Grand Challenge, driving autonomously for 131 miles along a desert trail it had never seen before. These successes were not due to some revolutionary new paradigm, but rather the tedious application of engineering skill and the tremendous increase in the speed and capacity of computers (Moore's Law).

In truth a new paradigm did emerge during the 1990s: Intelligent Agents (McCorduck 2004, pp. 471–478). Intelligent Agents are defined as any system that is capable of perceiving the environment and taking action maximizing its chances of success (Russell, 2010). The definition technically includes both humans and organizations of humans, but it characterized the AI field as the study of intelligent agents. This new approach allowed researchers to focus on individual problems and discover solutions that were both verifiable and practical. It provided a common language to describe problems and share solutions, both within and without the AI field.

This allowed researchers to come into contact with other fields such as mathematics and electrical engineering, an interaction that proved to be immensely beneficial for AI, both for its scientific rigor and for the inclusion of mathematical concepts in the field. These concepts, such as Bayesian networks, information theory, stochastic modeling and many more, constitute the very foundation of what is considered AI nowadays.

Around this time a rather interesting pattern started to emerge. As AI started building solutions to very difficult problems, these solutions (such as data mining, medical diagnosis, speech recognition, search engines) were being downgraded as simple computer science tools, rather than AI-powered tools. As Bostrom said (Bostrom, N. (2006), "AI Set to Exceed Human Brain Power", interview for CNN): "A lot of cutting edge AI has filtered into general applications, often without being called AI because once something becomes useful enough and common enough it's not labeled AI anymore."

This lead many researchers to deliberately call their work by other names, hiding their affiliation with the AI field in order to gain fundings from the cynic commercial world that still remembered AI winters.

Modern Age

The vast size of data that started to be collected with the popularization of the internet pushed AI to develop in the data analysis field. A large collection of information is today defined as Big Data and is characterized by the fact that it cannot be collected, managed, and analyzed by conventional software tools within a certain time frame. It requires a massive amount of decision-making, insight, and process optimization capabilities. The expansion in the Big Data field was possible thanks to cheaper and more powerful computers, as well as advanced machine learning techniques. Machine learning is a field of AI where computer systems learn from data, make predictions, and improve their performance without explicit programming. One of these advanced machine learning techniques is Deep Learning, a method that has become more and more popular. It leverages a deep graph with multiple processing layers to extract high-level abstractions from the data. In the present day, state-of-the-art deep neural network architectures have reached a point where they can occasionally match or even surpass human accuracy in various domains, such as computer vision. Notably, deep neural network have achieved impressive results in tasks like recognizing handwritten digits in the MNIST database or identifying traffic signs (Ciregan, Meier and Schmidhuber, 2012).

On the opposite side of highly specialized deep neural network we find general intelligence. General intelligence refers to the ability to solve any problem, rather than finding a solution to a specific set of problems. Artificial General Intelligence (AGI) is a program which can apply intelligence to a wide range of problems in much the same way humans can. It is also sometimes referred to as “Strong AI” and it has not been achieved yet, although lately there have been some important developments.

One of these developments is the creation of foundation models in 2018. These are expansive artificial intelligence models trained on extensive amounts of unlabeled data and can be fine-tuned to address a wide array of specific tasks. Significant milestones on the journey towards artificial general intelligence have been marked by the release of models such as GPT-3 by OpenAI in 2020 and Gato by DeepMind in 2022. These models have garnered recognition for their contributions in pushing the boundaries of AI capabilities and representing notable advancements in the pursuit of artificial general intelligence.

During 2023, Microsoft Research conducted extensive testing of the GPT-4 large language model across a diverse range of tasks. Their findings led them to conclude that GPT-4 could be considered as an initial iteration, though not yet a fully realized one, of an artificial general intelligence system (Bubeck et al., 2023).

1.2 OK Computer: AI in Music

The development of Artificial Intelligence and its application in the field of art seem to proceed hand in hand. Ever since music was formalized in mathematical terms by Pythagoras, the possibility of “producing” music through an automatic process was present. We can identify many attempts to rely on generative methods in order to create some form of music long before the invention of computers.

Aleatoric music dates back to as early as 15th century, one of the earliest examples would be the “Missa cuiusvis toni”, composed by Johannes Ockeghem. In this peculiar composition the singer decides the clef or key signature of the piece, purposefully left blank by the composer (Taruskin, 2010). A more famous example is the Musical Dice Game, even used by Mozart. In the game, a large number of musical measures are put in random order through the use of dices. Another, more modern example is John Cage’s 1951 “Music of Changes”, a composition for piano that was written by taking decisions through the I-Ching, a Chinese classic text commonly used as a divination system.

The first known application of computers in music comes from Alan Turing himself. In 1951, using the BBC outside-broadcast unit at the Computing Machine Laboratory in Manchester, Turing managed to make the computer “sing” three popular melodies (Copeland and Long, 2017). Following Turing’s attempt, five years later came Illiac I. Engineered by Lejaren Hiller, Illiac I composed many pieces of music, most notably “Illiac Suite String Quartet, No. 4”, which is considered to be the first substantial piece of music composed by a computer (Sandred, Laurson and Kuuskankare, 2009). Hiller achieved this goal through the use of Markov Chains, a probability method where the music is based only on the note that directly precedes it. In 1960 the first paper revolving around algorithmic music composition was published by Rudolf Zaripov (Zaripov, 1960). Shortly after, R. Kurzweil was able to create a computer capable of recognising patterns in various compositions. The system was then able to analyze and use the patterns to create new melodies. In 1997 David Cope created the “Experiment in Musical Intelligence” (or EMI), a system whose primary goal was to analyze a score and create variations in order to help with the composition process. Before long however it was able to replicate the intricacies of classical composers.

The advent of digital music production tools and the availability of large music datasets further propelled the progress of AI in music. These advancements opened up avenues for analyzing vast amounts of musical data, including audio recordings, sheet music, and metadata. Machine learning algorithms could now extract meaningful patterns, harmonies, and structures from these datasets, facilitating the creation of AI models capable of composing music. In recent years, deep learning techniques have revolutionized AI in music. Recurrent neural networks (RNNs) and generative adversarial networks (GANs) have played a significant role in advancing the field. RNNs, with their ability to model temporal dependencies, have been used to generate coherent and melodic sequences of notes. GANs, on the other hand, have demonstrated impressive capabilities in generating realistic and expressive music by setting a generator network against a discriminator network.

Today, AI in music encompasses a wide range of applications. AI models can compose original music, imitate the styles of specific composers, or even combine multiple musical genres to create unique fusions. AI systems can assist musicians in generating ideas, exploring different musical variations, and overcoming creative blocks. Furthermore, AI is being used to enhance music production, aiding in tasks such as audio synthesis, sound design, and automatic mixing and mastering.

1.3 Sympathy for the Scholar: A retrospective of AI Art Research

As stated, there is not much research on the sheer recognizability of AI music. As a matter of fact, recognizability is often studied in relationship with the subjective evaluation of the piece. There seems to be a known tendency of participants to report lower enjoyment when they are aware that the music they are listening to was created by an AI system. This bias seems to be confirmed by some researches (Shank et al., 2022; Moffat and Kelly, 2006) and confuted by others (Moura and Maw, 2021; Pasquier et al., 2016). A possible explanation is suggested by Hong, Peng and Williams (2021). In their research they found a significative relationship between the acceptance of the AI system as a musician and higher musical quality ratings. This suggests that the bias might be more present in people that refuse the idea of an AI musician and might have nothing to do with the music itself.

A rather interesting finding by Shank et al. (2022) is that listeners were more likely to attribute electronic music to AI rather than humans. This effect will be examined in the current paper.

Apart from the actual listening experience, people show a general skepticisim towards the current state of AI music, as shown by Knotts and Collins (2020) and low purchase intention (Moura and Maw, 2021).

During the creation and listening of AI music, Chu et al. (2022) found the most effective criterion to be melodiousness, with naturalness being important as well. Furthermore, they found that, when listening to AI-generated music, people value familiarity, emotion and replayability the most.

1.4 All Along the Watchtower: The Aims of this Research

This research comes from a deep curiosity towards AI generated media. Ever since the latest wave of interest in the matter, there has been much talk about the ethics and enjoyment of AI art, but not much about our ability to effectively discriminate between ”our” art and ”its” art. This research hopes to shed some standardized and formal light on the matter, providing an intuitive method to both frame and analyse the situation.

The primary hypothesis of this research is:

Hypothesis 1: Listeners can successfully discriminate between AI and Human music.

In the process of creating the task however, some more hypotheses were brought up:

Hypothesis 2: A higher score of music sophisticatedness correlates to a higher sensibility in discriminating AI and Human tracks.

Hypothesis 3: A higher level of familiarity with a given musical genre correlates to a higher sensibility in discriminating AI and Human tracks in that genre.

2 Method

2.1 Participants

A total of 72 participants (29 Male, 40 Females, 3 Non-Binary/Other; mean age= 21.65, sd = 1.92) have been tested. Participants have not been selected in any (conscious) way and have been reached employing multiple methods (flyers, word of mouth, presentations, messages in public group chats). The vast majority of participants were psychology students at the University of Padova.

All participants offered to take part in the study voluntarily and no compensation was offered. Data were collected from February to April 2023.

2.2 Stimuli and Setup

The study consisted of two main phases: the recognition task and the questionnaires.

The Recognition Task

The recognition task was created using PsychoPy and featured 30 audio tracks. Of these 30 audio tracks, 15 were AI-generated and 15 were human made. The audio tracks lasted 18 seconds each and were extracted from random points of the original whole songs.

The audio tracks could be further subdivided into musical genres. The genre label has been used in a broader sense at an instrumentation value. Each genre was presented an equal number of times (3 times as AI and 3 times as human). The genres have been selected arbitrarily and are: Jazz Trio (drums, upright bass, piano), Solo Piano, Rock (electric rhythm guitar, electric solo guitar, electric bass, drums), Electronic Dance Music or EDM, and Orchestral.

AI tracks have been generated with AIVA, one of the more popular and flexible music generators. AIVA is based on an algorithm that uses both machine learning and reinforcement learning architectures. At first the study was designed to feature multiple music generators, but AIVA was the only one to offer control both of the instruments and of the genre.

Human tracks have been selected both from personal knowledge and from the last positions of old song charts in order to avoid popular tracks. The exact tracklist (with timestamps for the sample used) can be found in the Appendix of this paper.
The tracks were presented in random order and, at the end of each presentation, the participant was asked to indicate whether they thought the track they listened to was human made or AI-generated. A third option was offered in case participants already knew the track.

The Questionnaires

After the recognition task, participants were asked to answer two brief questionnaires. The first one was the italian translation of the Goldsmith Music Sophistication Index (GOLD-MSI). In addition to the 49 items of the GOLD-MSI, two more items were added to the top regarding the participant's age and gender.
Following the Gold-MSI, participants were asked to answer a short questionnaire regarding their familiarity with some music genres. The answers were given on a 7 points Likert scale ranging from ”Not familiar at all” to ”Extremely familiar” and investigated the familiarity with the following music genres: Rock, Orchestral, EDM, Trap, Classical, Solo Piano, Jazz.

Setup

The laboratory contained an acoustically isolated chamber. Outside the chamber there were a desk, the computer, the monitor and a Focusrite Scarlett 2i2. The computer monitor was angled toward a glass panel mounted on the chamber. Inside the chamber were a chair, a desk with a keyboard and a mouse. The desk was positioned in front of the glass panel in such a way that the monitor was comfortably visible once seated. A headset was hanged on a plastic hook on the wall.

2.3 Procedure

Participants were welcomed into the laboratory and asked to fill in an informed consent. After having read and signed the document, the participants were brought into the acoustically isolated chamber. Before the recognition task, an explanation was given to the participants, featuring the aim of the research and the instructions for the task. The total number of audio tracks was omitted (but was reported on the informed consent). The proportion of AI and Human tracks was omitted as well to avoid participant focusing on the number of answers in each category. Once the task was started, the participants were left alone in the chamber, which remained closed for the entire duration of the task. The participant were told to knock on the glass panel once the task had ended, so that the researcher could enter and open the two questionnaires.

Once the questionnaires were completed, the experiment was over and participant could know, if they wanted, how many tracks they had successfully recognized.

For the sake of transparency, it must be stated that the questionannaire regarding musical genre familiarity has been introduced on the fourth day of data collection. In order to retrieve the answers from the first participants, a link for the questionnaire was sent to them. Considering the short length of the questionnaire and the intuitive stability of musical genre familiarity, this fact has not been valued as a threat to the validity of the answers.

Furthermore, before the genre familiarity questionnaire was introduced, the experiment came after the GOLD-MSI. When the familiarity questionnaire was added, the recognition task was moved as a way to avoid contaminating the data with the answers to the questionnaires.

2.4 Design

Since only one stimulus was presented at a time, the study conforms to an A-Not A design (Bi, 2015). Since more than one stimulus was presented, the design can be considered replicated.

Each trial can be considered as a Signal Detection test and evaluated in terms of Hit, Miss, Correct Rejection or False Alarm. From the data of each subject (mainly from the proportion of hits and false alarms), a sensibility index (*d'* or *d prime*) and a criterion index (*c*) can be calculated and used in the analysis.

The third option of the task (to be chosen in case the participants already knew the track) was originally meant to be interpreted as an invalid trial. During the analysis however, for the sake of not overcomplicating the computations, it was treated as a ”Human” answer. This means that participants were given the benefit of the doubt regarding their real knowledge of the track presented, probably leading to a light overestimation of the actual sensibility.

3 Results

General Performance Anaysis

At the end of the data collection, a mean accuracy of 62.8% was found. This percentage includes both the accuracy for human music recognition (that stands alone at 69.6%) and for AI music recognition (56.2%).

Considering this as a simple binary choice would require us to set the required percentage at 75% so that it stands in between the random choice scenario (50%) and the completely correct scenario (100%). This framework suggests the conclusion that participants were not able to correctly discriminate between AI and Human music.

However, the mere analysis of the percentage constitutes a superficial approach that does not take into account parameters such as the tendency of subjects to answer AI or Human more often than not. In order to analyze the data in a more appropriate approach, the framework of Signal Detection Theory (SDT) is the best fit.

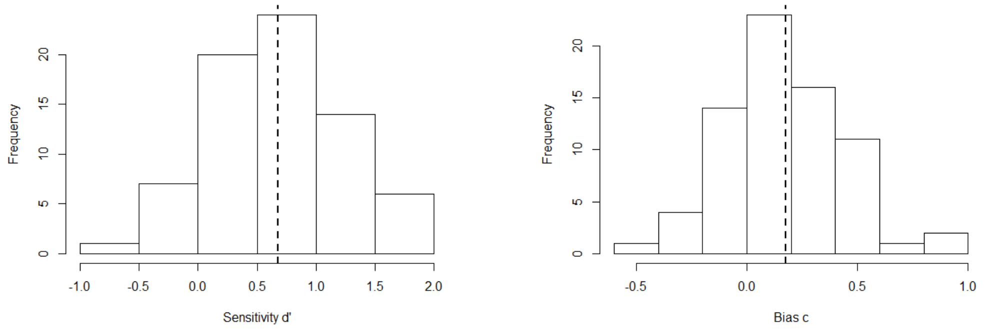
Before proceeding with the analysis, a Chi-Square Test of Independence was performed to assess the relationship between stimulus presented and answer given. The null hypotheses is that the two variables are independent, meaning that participants answered in a random way. The test proved significative, allowing us to reject the null hypotheses and claim that answers were not given randomly (X^2(1, N=2160) = 146.80, *(*p) = $<$ 0.00001).

Framing each answer as either a Hit (answer is AI when track is AI), a Miss (answer is Hu when track is AI), a False Alarm (answer is AI when track is Hu) or a Correct Rejection (answer is Hu when track is Hu), we can calculate two main values for each subject. The first one is called *d'* (d prime) and is calculated by substracting the normalized proportion of Hits and False Alarms. *d'* indicates how accurately the subject was able to detect AI music. The second index is *c* (which stands for criterion) and is calculated by adding the normalized proportion of Hits and False alarms and then dividing that value by two. *c* indicates whether the subject had a bias for answering AI or Human more often than not.

In order to avoid problematic values during the calculations, the formula adjustments from H. Stanislaw and N. Todorov (1999) were used. Descriptive analysis for SDT parameters, *d'* and *c* can be found in Table 1. The distribution of *d'* and *c* values across the whole sample can be found on Figure 1.1. Furthermore, the complete analysis of errors and musical genres can be found in Table 2.

	Mean	**Standard Deviation**	**Min.**	**Max.**
Hit rate	0.558	0.133	0.281	0.906
False Alarms rate	0.316	0.116	0.031	0.594
Sensitivity *d'*	0.675	0.530	-0.658	1.941
Response bias *c*	0.176	0.262	-0.540	0.892

Table 1: Descriptive analysis of Signal Detection Theory.



(a)

(b)

Figure 1: Histograms for sensitivity (Figure 1a) and response bias (Figure 1b) values.

Musical Genre	**Miss (track is AI, answer is Hu)**	**False Alarms (track is Hu, answer is AI)**	**Track is AI, answer is ”Known”**	**Total**
EDM	84	120	0	204
Solo Piano	126	63	0	189
Orchestral	135	37	2	172
Rock	63	78	3	141
Jazz	63	30	1	93

Table 2: Errors per Musical Genres.

Musical Sophisticatedness and Performance

As stated by Hypotheses 2, a positive relationship between Gold-MSI scores and sensibility was expected. This relationship has been investigated through the use of Pearson’s r, considering the five different sub-scales and the General Sophistication scale. Results can be found in Table 3.

GOLD-MSI	Pearson’s *r*	*p* (two-tailed)
Active Engagement (F1)	0.081	0.500
Perceptual Abilities (F2)	-0.030	0.805
Musical Training (F3)	0.016	0.895
Emotions (F4)	0.027	0.821
Singing Abilities (F5)	-0.149	0.211
General Sophistication (FG)	-0.038	0.748

Table 3: Correlations of Gold-MSI and sensitivity d’.

Musical Genre Familiarity and Performance

As stated by Hypotheses 3, a positive relationship between familiarity with a music genre and the ability to recognize AI music of that genre was expected. This relationship has been investigated through the use of Pearson's r, calculating the relationship between the proportion of correct answer of one genre and the familiarity score given by the participant. Results can be found in Table 4.

Musical Genre	Pearson's *r*	*p* (two-tailed)
EDM	-0.159	0.182
Jazz	0.137	0.250
Orchestral	0.004	0.976
Piano Solo	-0.091	0.448
Rock	0.154	0.200

Table 4: Correlations of Musical Genre and correct answers for that genre.

Notes

Following the experimental session, while escorting participants back to the entrance, there were frequent observations regarding the strategies employed during the recognition task. Initially raised by the first participants, it soon became a habit of mine to investigate these strategies. Although not in a standardized manner, I began documenting the more distinctive strategies employed. Among the most prevalent strategies, assessing the emotional impact of the music emerged as the foremost approach, with the implicit notion that pieces lacking sufficient evocative or expressive qualities were generated by AI. Another commonly used strategy pertained to the perception of ”coherence” in the piece, encompassing both the dynamic aspect related to harmonic structure and the sonic aspect related to the integration of different sounds. The third most prevalent strategy involved seeking audio cues indicative of the physicality of the instruments, such as the piano's pedal or the sound of fingers on the keys. Multiple participants explained that in their attempts to determine the author, they relied on visualizing a human musician performing the exact piece on the instrument, and their judgment was based on the perceived plausibility of this mental image. An intriguing yet infrequent strategy involved focusing on the application of silences, where pieces that incorporated purposeful pauses were deemed to be of human origin.

Due to the absence of standardized data collection, obtaining statistical evidence for the significance of these strategies and their relationship to performance proves challenging. Therefore, only through a superficial analysis, no apparent significant relationship seems to exist between the strategies employed and the participants' performance, framing this as nothing more than an interesting consideration.

4 Discussion

As far as Hypotheses 1 is concerned, a mean sensibility of d’ = 0.675 resides in the range of ”rather small difference” (d’ between 0.0 and 0.74) (Bi, 2015). This suggests a performance clearly above the chance level but not high enough to be considered meaningful (d’ = 0.74 and above). Of the whole sample, only 28 participants (38.9%) exhibited a sensibility value above 0.74. In the end, Hypotheses 1 has been confuted.

Hypotheses 2 and 3 have also been confuted, with a failure to prove that one’s musical sophisticatedness or familiarity with a given music genre act as an influence in the ability to discriminate AI tracks in that genre.

Additionally, the bias previously shown by Shank et al. (2022) by which listeners tended to attribute electronic music to AI has been confirmed since EDM is, overall, the genre with the most False Alarms.

5 Conclusion

In this study we used an A-Not-A design to test whether participants were able to successfully discriminate AI music from Human music. Results clearly indicate that, even if performance was different from random choice, the sensitivity is not high enough to affirm that participants were ultimately able to distinguish the two types of music.

Furthermore, no significant relationship was found between musical genre familiarity and ability to discriminate AI music of that particular genre.

In addition to musical genre familiarity, musical sophisticatedness was also tested. Results indicate no meaningful relationship between musical sophisticatedness and performance.

6 References

Bi, J. (2015). *Sensory discrimination tests and measurements: Sensometrics in sensory evaluation* (2nd ed.). John Wiley & Sons.

Brooks, R. A. (1990). Elephants don't play chess. *Robotics and Autonomous Systems, 6(1-2)*, 3-15.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv:2303.12712v5.*

Castro, S. L., and Lima, C. F. (2014). Phase synchrony analysis of EEG during music perception reveals changes in functional connectivity due to musical expertise. *Music Perception, 32*, 125–142.

Chu, H., Kim, J., Kim, S., Lim, H., Lee, H., Jin, S., Lee, J., Kim, T. and Ko, S. (2022). An Empirical Study on How People Perceive AI-generated Music. *In Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 304-314.

Ciregan, D., Meier, U., Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. *Conference on Computer Vision and Pattern Recognition, 2012*, 3642-3649.

Copeland, B.J., Long, J. (2017). Turing and the History of Computer Music. In: Floyd, J., Bokulich, A. (eds) *Philosophical Explorations of the Legacy of Alan Turing,* Boston Studies in the Philosophy and History of Science, vol 324. Springer Cham.

Crevier, Daniel (1993). *AI: The Tumultuous Search for Artificial Intelligence*. New York: BasicBooks.

Crossley, J. N. (2005). Raymond Llull's Contributions to Computer Science. *Monash University.*

Feigenbaum, E. A., Buchanan, B. G. (1993). DENDRAL and Meta-DENDRAL roots of knowledge systems and expert system applications, *Artificial Intelligence, 59,* 233–240.

Fidora, A.M., Sierra, C., Barberà, S., Beuchot, M., Bonet, E., Bonner, A.J., Colomer, J.M., Crossley, J., Sales, T., and Wyllie, G. (2011). *Ramon Llull: From the Ars Magna to Artificial Intelligence*. Barcelona: Artificial Intelligence Research Institute, IIIA.

Hong, J. W., Peng, Q., and Williams, D. (2021). Are you ready for artificial Mozart and Skrillex? An experiment testing expectancy violation theory and AI music. *New Media & Society, 23(7)*, 1920-1935.

Knotts, S., and Collins, N. (2020). A survey on the uptake of Music AI Software. *Proceedings of the International Conference on New Interfaces for Musical Expression*, 499-504.

McCorduck, P. (2004). *Machines Who Think.* Natick: A K Peters, Ltd.

McCulloch, W. S., Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115–133.

Moffat, D. and Kelly, M. (2006). An investigation into people's bias against computational creativity in music composition. *Proceedings of the third joint workshop on Computational Creativity.*

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine, 27(4)*, 87.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhard, C., Forghani, B., Forghani, R. (2020). Brief History of Artificial Intelligence, *Neuroimaging Clinics of North America, 30 (4)*, 393-399.

Newquist, H. P. (1994), *The Brain Makers: Genius, Ego, And Greed in the Quest For Machines That Think.* New York: Macmillan/Sams Publishing.

Pasquier, P., Burnett, A., Thomas, N. G., Maxwell, J. B., Eigenfeldt, A., and Loughin, T. (2016). Investigating listener bias against musical metacreativity. *In Proceedings of the Seventh International Conference on Computational Creativity*, 42-51.

Russell, S. J., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Pearson Education.

Sandred, O., Laurson, M., Kuuskankare, M. (2009). Revisiting the Illiac Suite–a rule-based approach to stochastic processes. *Sonic Ideas/Ideas Sonicas, 2*, 42-46.

Schaeffer, J., Plaat, A. (1997). Kasparov versus Deep Blue: the rematch. *J Int*

Comput Games Assoc., 20, 95-101.

Shank, D. B., Stefanik, C., Stuhlsatz, C., Kacirek, K., Belfi, A. M. AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied, 2022 Aug 25.*

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Simon, H. A. (1965). *The Shape of Automation for Men and Management.* New York: Harper & Row.

Stanislaw, H., Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers 31*, 137–149.

Taruskin, R. (2010). *Music from the Earliest Notations to the Sixteenth Century. The Oxford History of Western Music. Vol. 1.* New York: Oxford University Press. 479–480.

Tigre Moura, F., and Maw, C. (2021). Artificial intelligence became Beethoven: How do listeners and music professionals perceive artificially composed music? *Journal of Consumer Marketing, 38(2),* 137-146.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, s2-42,* 230-265.

Turing, A.M. (1950). Computing Machinery and Intelligence, *Mind, LIX (236)*, 433-460

Zaripov, R. (1960). On algorithmic description of process of music composition. *Proceedings of the USSR Academy of Sciences, 132.*

.

Appendix

Song	**Artist(s)**	**Name of the file**	**Timestamp**
Danny's Blues	Ennio Morricone	HU_P1	1:48-2:06
Introspection	Thelonious Monk	HU_P2	1:30-1:48
Playing Love (Piano Version)	Ennio Morricone	HU_P3	2:24-2:42
Orchestral Suite (Moon River), Breakfast at Tiffays	Henry Mancini and Johnny Mercer	HU_O1	3:54-4:12
Downtown Abbey (Main Theme)	John Lunn	HU_O2	2:42-3:00
Extract from the Centennial Concert of Field of Dreams	James Horner	HU_O3	7:47-8:05
Peaches and Diesel	Eric Clapton	HU_R1	0:54-1:12
Son of Alerik	Deep Purple	HU_R2	9:00-9:18
Coast to Coast	Scorpions	HU_R3	4:12-4:30
How Deep is the Ocean	Bill Evans Trio	HU_J1	2:42-3:00
The Pond	Jeff Hamilton Trio	HU_J2	3:18-3:36
Autumn Leaves	Beegie Adair Trio	HU_J3	4:30-4:48
Magnetic Fields	Tomas Skyldeberg	HU_E1	3:18-3:36
Lucy's Game	Senchi	HU_E2	0:36-0:54
Crystal Kid 3	Tomas Skyldeberg	HU_E3	4:48-5:06

Tracklist.