Explainable Artificial Intelligence And User Experience

Measuring the Effectiveness of a Communication Interface

Department of Computing

Master's Thesis

Villeveikko Sula

December 2024

Master's Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Artificial Intelligence
Author: Villeveikko Sula
Title: Explainable Artificial Intelligence And User Experience: Measuring the Effectiveness of a Communication Interface
Number of pages: 57 pages, 29 appendix pages
Date: December 2024

A recent focus in the field of artificial intelligence (AI) has been revolving around generative AI and deep neural networks utilized by them. This technology is often opaque by its nature, with AIs utilizing them often referred to as *black boxes*. Explainable artificial intelligence (xAI) is being actively researched, in an effort to add transparency to the AI algorithms. However, different kinds of xAI methods can produce immense amounts of data, which may not help a normal user in any way.

This thesis presents a chat-based communication interface attached to an AI framework made for the videogame *Super Mario Bros.* The chat interface was used for transferring information from the AI agents playing the game to the user, as well as for a chatbot reacting to the game state. A survey was made to find out how the chat features affected the way the user or observer evaluated the intelligence of the AI. A total of 65 answers were collected.

The results show that including the chat features lowered the average confidence of the responders, especially the ones who were very familiar with *Super Mario Bros.* The inclusion of the chat interface did not seem to affect the evaluations by much, otherwise. Importantly, the more time the responder spent on reviewing the AI and the chat interface, the more accurate their evaluations started to be, as the AIs presented in the survey were fake.

Keywords: Explainable Artificial Intelligence, Communication Interface, User Experience

Table of contents

1 Introduction

The field of artificial intelligence, AI, has been developing since the 1950s. While we are probably nowhere close to creating something truly intelligent, many developed AI techniques and technologies have seen practical use over the years. For example, generative AI has been popular in recent years, with ChatGPT surprising everyone with its text generation capabilities (Open AI, 2022), Midjourney generating pictures from a text prompt (Feingold, 2022), and Sora for generating video from a text prompt (Brooks et al., 2024). In the wake of the success of the generative AI models, many have proclaimed that AI development has finally made a breakthrough and that it is only a matter of time until almost everything will be solved by AI.

However, this is not the first time when a similar excitement emerged in artificial intelligence research. It is perhaps telling that this is something that was already discussed in the 1990s: For example, one paper coined the term *gee whiz view*, which "maintains that for a particular task, if no machine ever did it before, it must be AI." (Schank, 1991) While the term itself is quite tongue-in-cheek, it highlights that even back then it was recognized that one was easily blinded by good results, which could make a software appear more intelligent than it actually was. Surely if a machine can differentiate objects or text in an image, or do some other difficult task, it must be intelligent in one way or another. Conversely, this can also cause the software to not be considered AI anymore: After the effect has worn off and the software's features become commonplace, it consequently seems less intelligent (Ibid.).

On a more formal level in computer sciences, the discussion regarding artificial intelligence revolves often around an agent's ability to solve various logical problems in a multitude of environments (Legg & Hutter, 2007; Hernández-Orallo & Dowe, 2010; Gao et al., 2019). This works as a definition for artificial or machine intelligence — and one could argue it works as a definition for intelligence as well. These kinds of formal definitions are useful for the field because they allow for formal testing. It also helps mitigate the gee whiz view, as human impressions are replaced by a formal method of calculating intelligence. It also enables AI researchers to say that an algorithm or a program is more intelligent than another.

But no matter how one defines artificial intelligence and the tests to measure it, there exists a challenge to produce reliable results. This is called the *Clever Hans effect*. Clever Hans refers to a horse that lived from the late 19th to the early 20th century that could solve various arithmetical problems presented by its owner. Later it was found that the horse produced reliable results not because it could actually count, but because it could read its owner's subconscious cues so well that it could tell what the right answer was based on that (Samek & Müller, 2019). This effect is a very prevalent issue in the field of AI research, as we need to know that the AI produces reliable results for the right reasons, and not because of how the test is defined or made.

What makes this even more pressing is the fact that many popular learning algorithms do not symbolize their learned representations (Samek & Müller, 2019). That is, it is nigh impossible to observe what the algorithm has learned and why, so we can only speculate about the kind of deductions the AI makes from its learning material (Ibid.). The deductions that the AI makes can be incorrect, even if the end result would be correct. A fitting example of this is when an image recognition algorithm learned to tell whether there was a horse in a picture based on the watermark in the corner of the image (Lapuschkin et al., 2019), or when an algorithm differentiated a "husky" from a "wolf" based on whether there was snow in the picture (Ribeiro et al., 2016).

Phenomena like the gee whiz view and the Clever Hans effect are important to notice because despite numerous advances, they remain as one of the big challenges in artificial intelligence research. From the historical observations of Schank (1991) we can and should conclude that it is not enough for AI to just be able to do something new. Additionally from the various investigations of Samek & Müller (2019), Lapuschkin et al. (2019), and Ribeiro et al. (2016), we know that measuring the intelligence of AI solely based on the tasks that it does is unreliable. With this information it can be claimed that any AI that does not answer these challenges may not be considered AI in the future.

All of this helps explain the ups and downs of artificial intelligence research over the years: The rapid development from the 1950s to the 1960s (Newquist, 1994), the *AI winters* around the 1970s and the 1990s (Ibid.), and now an increase in funding and investment from 2012 to this day (Lohr, 2016). And while AI research is still being invested in, if those investments do

not bear results, investors in the field might become more wary. It is therefore possible that eventually we could be heading towards a new AI winter, with increasing promises on what the AI is capable of doing, but having various problems when putting them into practice[1]. When something goes wrong and the intelligence of the AI is questioned, it will eat away at the trust towards AI projects and reduce the available funding in the future.

In essence, both the gee whiz view and the Clever Hans effect are caused by unclear communication. It does not help that we as humans have a tendency of anthropomorphizing other things, whether they are animals or machines (Kim & Sundar, 2012). An observer may be naturally inclined to think that because a program or an animal ended up with the same result as the observer themselves would with the given problem, it must have got to that result with the same or similar reasoning. Verifying the animal's or the program's actual reasoning might be very difficult or impossible for the observer. The same reason can also cause them to misjudge an AI's capabilities.

It is worth noting that the opposite is likewise possible: If the observed thing does not produce results that we understand, it is often dismissed as less intelligent and incapable of doing the task. This is also caused by unclear communication. For example, it was found out that cats were not worse at following instructions than dogs, they were just less interested when a stranger spoke to them (de Mouzon et al., 2023). In a similar manner fish were thought of as incapable of feeling pain or stress just because they did not express that feeling the same way many mammals do (Braithwaite & Ebbesson, 2014). Likewise bees were thought of as incapable of learning, until it was found out that they were capable of solving puzzles made by researchers by watching other bees solve them (Bridges et al., 2023).

In a similar way, it is incredibly easy to misunderstand AI just by looking at the results it produces. An important distinction is that we are not restricted by a language barrier with artificial intelligence. The output that a program produces and the input that it takes in are within our power to define and modify, and as such, it is possible for a program to produce more than just the end result. It would be within our interests for AI development to have as much information as possible about what the program is doing and why. This could be called a form of communication, or a transparency protocol, and it would still be a challenge to

[1] For example, self-driving cars have been pushed for almost a decade now as a safer option to manual driving, but recent accidents have not won the public opinion over. See (Zhang et al., 2024).

program. But when implemented with AI, it has been speculated that it would reduce at least some of the obfuscation present in many AI implementations (Samek & Müller, 2019).

Communication in intelligence is given varying degrees of importance, but it has been there since the beginning. The Turing Test (Turing, 1950) is a prime example of this, as its focus is solely on communication. Though as time has passed and the Turing Test has received its share of criticism (ACM, 1992), AI development has focused more on the results. One could say that even with the Turing Test, the goal of AI has always been some kind of an end result, whether it is forming a sentence or moving a chess piece on the board.

This can be seen in artificial intelligences that specialize in communicating with people. In recent years, there have been instances where people have called an AI sentient because of how well it communicated with its user (De Cosmo, 2022), causing discussion about socially manipulative intelligence (Carroll et al., 2023). While using machine learning algorithms produces very interesting results, they fall under the same problem as our previous examples. The usage of neural networks and similar techniques obfuscate the program's process of producing sentences, which again produces more questions than answers. For example, when ChatGPT works as expected, the general public do not tend to question how it works, but when it does not, they start wondering about the underlying issues and OpenAI's vague answers about them (Stokel-Walker, 2024). Communicative programs are not really that different from self-driving cars, producing impressive results but leaving the observer completely oblivious of their way of operation.

The obfuscated nature of modern AI solutions and the possible methods to add explainability and transparency to them motivated me to do this thesis. It is my belief that we should consider communication in AI as a part of the solution; not as something that one can ignore, but also not as an end result. Communication works as a proof of intelligence, hence it is essential for the artificial counterpart as well. But as we have seen, communication in programs is not a trivial matter. It can be difficult to say what level of communication is needed to prove the inner workings of AI adequately. All solutions might not need an introspection-level of communication, but for some other, high-risk tasks, implementing such features might be crucial in order to trust an AI enough to do them.

There are some existing solutions for explainable artificial intelligence, also known as xAI (Saranya & Subhashini, 2023). These solutions answer the question of getting information about how the AI works, but they are not really concerned about presenting such information. Having the information is only the beginning of successful communication, and an AI developer still has to decide how to present the data: Is there so much data that it needs to be filtered? Should the data be shown only when asked, or can the program show it immediately when it is relevant? What should the UI for this communication look like? Is the communication one-way, or should the user be able to affect it? These questions are the ones that interest me the most, and there does not exist a lot of research about it, so it will be the focus of my thesis.

I set out to design a communication prototype on top of an artificial intelligence platform. The result of that investigation was a chat user interface added on top of an open source *Super Mario Bros.* -videogame environment called Mario AI Framework, made by Khalifa (2019), as well as an interface that the AI agents playing the game in that environment could use to provide information to the chat. The user could type various sentences into the chat to essentially communicate with the AI agent, issuing it rudimentary commands. Additionally, the chat itself would archive the decision-making of the AI in a readable format, enabling the user to query such information afterwards. The system was designed to work with different kinds of AIs that implemented the interface with the chat, adding a layer of transparency to all of them. Technical review of this project is provided in Appendix A.

Besides the possible benefits of transparent communication in AI, I was also interested to see if there were some unintended consequences because of it for the human-AI interaction. For example, does simplistic but transparent communication make every AI appear stupid to the observer? Comparing different kinds of communication, the quantity of the messages, the amount of information in the messages, the agency of the user and how they help or hinder the observer would provide valuable insight when considering communication between AI and the user in future applications.

For this reason, the objective in this thesis is to look into how my solution to AI communication affects an observer's perception of the AI, compared to when such communication does not exist. My hypothesis was that the added communication from the AI

improves the observer's ability to assess the intelligence of the AI, even if the level of communication is quite limited. This would not only mean that the observer gets more confident in their feeling of the AI's intelligence, but that the observer would also be able to tell when the reasoning of the AI is questionable or outright wrong.

The methodology for this thesis is survey research. The survey has observers of the AI evaluate its intelligence by rating various statements about it on a Likert scale[2]. By evaluating the AIs with and without the chat prototype, these quantitative answers can be compared and analyzed to see if the responders evaluate the AIs differently with different amounts of communication provided to them. Based on these results, it can be assessed if a chat-based communication has an effect on an observer in AI environments, and whether this effect is positive or negative.

This thesis will first introduce current popular algorithms and techniques used in artificial intelligence, and highlight their challenges in providing information on how they work. After that, existing ways of communication in AI, and what they provide, are considered. From this background, the structure of the communication interface is explained, and go through it in-depth. The survey I created for measuring the interface is then presented, with explanation of its structure and the methodology for evaluating the answers. Responses to the survey are then analyzed, and the implications of the results are discussed. It is also discussed what could have been done better for the survey in the thesis, as well as future points of work that would contribute to the subject.

[2] A Likert scale is a numerical rating scale that the survey responder uses to indicate how much they agree with a statement, with higher numbers indicating agreement and smaller numbers indicating disagreement. For the survey in this thesis, a five-point Likert scale from 1 to 5 was used. Instead of numbers, a Likert scale can use enumerations of those numbers, for example "fully agree" representing value 5, or "slightly disagree" representing value 2.

2 Popular AI Algorithms And Communication

Before evaluating the effectiveness of communication in artificial intelligence, it is important to understand how AI algorithms can provide information in the first place. One of the prevalent challenges in recent AI research has been to increase the transparency of its algorithms, such as artificial neural networks, and it would not be an exaggeration to say that there is still a lot of work to be done (Samek & Müller, 2019; Saranya & Subhashini, 2023). By examining some of the existing solutions it is possible to get a picture of how this kind of information could be received, which in turn helps in designing a communication interface between the AI and the user.

In artificial intelligence development, many algorithms have gotten their ideas from nature and biology (Basheer & Hajmeer, 2000; Ventura et al., 2022). It follows that some of the common themes in those algorithms include parallelism and nonlinearity, the ability to learn or adapt in some manner, and the ability to withstand noise[3] and uncertainty (Basheer & Hajmeer, 2000). This thesis goes through different AI algorithm techniques, based on recent papers about what type of AI algorithms are currently in use in practice (Onyelowe et al., 2023; Huang et al., 2019). These techniques and algorithms are Artificial Neural Networks (ANN), Genetic Algorithms (GA) and Fuzzy Logic (FL).

This chapter will briefly present and explain each algorithm and technique, and go through why they are effective. Then, the transparency of the algorithm or technique is examined, and it is discussed how difficult it would be to extract an explanation out of the algorithm or technique, if at all possible. Finally, an example project or program that utilizes the algorithm or technique will be given, highlighting its features.

2.1 Artificial Neural Networks (ANN)

Artificial neural networks are very popular algorithms in use in AI solutions. Their design is loosely based on biological neurons: The artificial neurons receive stimulus from the input, with one kind of input activating a set of neurons. This set can be then provided to another set

[3] Noise refers to data that is not relevant to the algorithm's task at hand, or to data that has been corrupted in one way or another.

of neurons as input, and this chain is continued until an output level is reached, where the set of activated neurons determine the output value. The learning process of these algorithms is about adjusting the weights of the connections between the neurons, meaning how easily the neuron activates to which stimuli, after reviewing whether the generated output matches with the desired output. The desired output can be determined with training material, or some other mechanism, such as a scoring system (Basheer & Hajmeer, 2000).

When implemented correctly, ANNs can become great at generalizing new data based on what the network has learned from the training data. For example, recognizing things from images, such as faces in a mobile phone's camera app, is often done with the help of ANNs (Samek & Müller, 2019). However, ANNs struggle with providing a reason why an output was generated from the given input:

> *Complex classifiers such as deep neural networks or recurrent models on the other hand contain several layers of non-linear transformations, which largely complicates the task of finding what exactly makes them arrive at their predictions.*
> (Samek & Müller, 2019)

Due to this phenomenon, these kinds of algorithms are often called *black box algorithms*, as the inner workings of the algorithms are almost completely opaque for the viewer (Ibid.). These problems with transparency become problems with communication, as the ANN has no way to defend its decision-making processes. This can make it difficult to trust these algorithms in situations where knowing the decision-making process is important, such as with health-related topics.

There has been an effort to make artificial neural networks more transparent in various ways. Some studies have tried to reveal what the neurons in the network represent, either by dissecting them or by making the network build a "prototype" of a category, for example, what the network thinks "a car" looks like. It is also possible to look into the individual predictions of the network by studying which aspects of the input are the most important for the network to reach to its conclusion. These individual predictions can also be used to explain the model behavior in general, for example by clustering individual heatmaps to create meta-explanations. Lastly, an alternative approach would be to look at the training

dataset and identify representative examples from it that explain the model behavior (Samek & Müller, 2019).

AlphaGo is by now a classic example of ANNs achieving better than human-level performance in a complex board game. Utilizing multiple ANNs and Monte Carlo tree search (Silver et al., 2016) for the game of Go, AlphaGo was able to achieve the efficiency and accuracy needed to beat both the 2 dan European champion Fan Hui (Ibid.), as well as the 9 dan professional South Korean player Lee Sedol (Samek & Müller, 2019). During the game against Lee Sedol, AlphaGo played a move that a Go expert classified as “not a human move”, and it was only found out later on why it had played that move to gain an advantage (Ibid.). This highlights the potential ANNs have to find out new solutions for problems, but also their opaque nature: Something that seems extremely weird at first can turn into a brilliancy later on, but having a mechanism to see into the reasoning of the algorithm in the moment would help us to trust the choices it makes.

2.2 Genetic Algorithms (GA)

Genetic algorithms search for an optimal way to reach a desired solution in a semi-random manner. The reason why they can be considered an artificial intelligence technique is because they are a type of evolutionary algorithm: A set of random solutions, called chromosomes, to a problem are evaluated, and the ones that work the best are then picked to be mutated or combined with each other. The result becomes the new set that is tested, and this cycle continues until a solution reaches the desired performance or accuracy in the test, meaning that an optimized solution is found. This solution can then be picked and used for any problem within the framework of the test (Ventura et al., 2022).

While the idea is relatively simple, genetic algorithms have proven to be effective for optimizing complex systems, while being general enough to not depend on the particular field of the problem (Huang et al., 2019). If done correctly, a solution utilizing a genetic algorithm can offer fast and accurate results while adhering to almost any kind of constraints given during development (Ventura et al., 2022). As long as the “test problem” during the development is defined broadly enough, genetic algorithms can also perform adequately in

unfamiliar situations. This can come at a cost of performance, however, when compared to a highly specialized algorithm (Ibid.).

Genetic algorithms do not have similar issues with transparency as artificial neural networks, as the chromosomes can generally be translated into an understandable solution (Ventura et al., 2022). However, they have very little to say about themselves when it comes to communicating what they do, as the solution to the problem has been found in a random manner. The only thing such an algorithm can answer when asked about the reason for its solution is something similar to "Because, through trial and error, it was found out that it took the least amount of time." It does tell the user that the algorithm has tried various things and settled with a solution, but it does not reveal potential mistakes in the solution.

Because of this, translating the found solution into an understandable format is an important step in making genetic algorithms more understandable and transparent. That in itself provides a way for a user to detect errors or insights in GA's programming, although it can be a rather laborious effort, since the solutions can be quite complex due to the complexity of the system. Powerful tools able to dissect and analyze the chromosomes should be utilized when relaying information to the user. Translatability should also be kept in mind when designing the structure of the chromosome, to make the process easier.

It is important to note that genetic algorithms are often not the whole AI solution, but used in conjunction with other AI techniques, such as the aforementioned artificial neural networks (Ventura et al., 2022). If this is the case, it is possible that GA inherits the transparency issues of the other algorithm. It is imperative that each component going into an AI solution is transparent and able to transfer understandable information to others, so that as much information as possible can be given to the user.

There are other artificial intelligence techniques, such as swarm intelligence, that are quite similar to genetic algorithms. Unlike GA, a swarm intelligence algorithm does not consider individual solutions as chromosomes, but as individuals. Each individual is affected by their neighbors and they are all seeking the optimal solution to the given problem. So if a neighbor has a better solution than an individual, the individual can fully or partly adapt the neighbor's solution. Eventually, most of the individuals have flocked towards the optimal solution, by

which point the solution can be picked and used (Kennedy et al., 2001). Swarm intelligence algorithms can be treated in a similar way as GAs in regards to their transparency.

Researchers (Wua et al., 2012) combined a support vector machine with genetic algorithms to identify whether a breast tumor was benign or malignant based on a single gray-scale image. A GA was used to first distinguish the importance of 30 different classified features when it comes to diagnosing the tumor. Then, using the set of features deemed important, another GA was used to fine-tune the parameters needed by the support vector machine algorithm, which could then be used to categorize the tumor images. The test results with real patient data showed that this approach yielded results with over 95% accuracy (Ibid.). This example illustrates how GAs can be used to optimize other AI algorithms.

2.3 Fuzzy Logic (FL)

Fuzzy logic algorithms have become common in the use of practical AI solutions where uncertain elements are often present (Onyelowe et al., 2023). Essentially, FL algorithms enable using performant logical operations with non-binary values. Typically, these fuzzy values are decimal numbers ranging from 0 to 1, where 0 refers to false and 1 to true. Fuzzy logic algorithms can determine an entity to partly belong to categories, which can be used to model real-life situations in a more realistic manner (Ibid.).

By accepting uncertain data for its operations, fuzzy logic has the benefit of being able to produce as accurate results as possible with limited information. Another side of FL is that it is more connected to our language, and as such it is more translatable: For example, a typical output of an FL algorithm could be translated into a percentage value, “There is a 71% chance that it is going to be sunny tomorrow”, but also to a fuzzy value, “It will probably be sunny tomorrow”. This applies to its parameters as well, which makes the algorithm often understandable (McNeill & Thro, 1994).

Many AI solutions often deal in estimates, rather than in absolute values. This makes fuzzy logic algorithms well suited for being paired up with other AI techniques. For example, artificial neural networks can be used for creating fuzzy rules, or for utilizing fuzzy control systems (McNeill & Thro, 1994). One of the better known contemporary implementations is called the Adaptive Neuro Fuzzy Inference System, also known as ANFIS. The idea of

ANFIS is to combine linguistic principles of FL with ANN, by fuzzifying the input for the ANN and then de-fuzzifying the output coming from the ANN (Onyelowe et al., 2023).

What makes ANFIS a very interesting concept for this thesis is that by introducing fuzzy concepts to ANNs, the network can potentially become translatable, as it is operating with the fuzzified data that can be translated. This has been researched in (Keneni et al., 2019), and it shows that getting a decision from ANN can be explained afterwards, as long as the network has been fuzzified. This has the potential to solve some of the transparency issues in ANNs that were discussed before. However, adding a fuzzy control system to a solution is not always feasible, and can require too much effort with more complex systems.

One of the earliest commercial applications of fuzzy logic goes all the way back to 1987, when the city of Sendai in Japan took the Sendai Subway 1000 series into use (McNeill & Thro, 1994). Fully electric and automated, the trains also used about 10% less energy when compared to human drivers (Bird, 2021). This can be largely attributed to the fuzzy logic system, where the acceleration and deceleration can be handled smoothly and changed on the fly when there are changes in the environment. While this is a simple and old example, it illustrates the core strength of using fuzzy logic in non-binary situations.

From the examples described in this chapter, it can be seen that while there are transparency challenges with some of the existing AI algorithms, it is possible to start solving them by using other AI techniques. For example, combining fuzzy logic with ANNs could potentially produce understandable explanations for specialized deep learning networks. And if it was possible to use a genetic algorithm to improve another one, it would not be a stretch to think that it would be possible to use ANNs against each other in a similar manner, in order to dissect and understand them.

There is clearly potential to find explainability in AI algorithms, but little has been discussed about how this explainability should be shown. It is one thing to store all the information about the AI's decision making process, but it only benefits the analysts and experts behind the AI. Simply showing all of it to the user is not sufficient either, as massive amounts of information would most probably just confuse them. How does the AI determine what information is wanted or needed? How is the excessive information detected and handled?

How do we increase the interactivity with the AI and the user? This thesis describes a possible solution for that in the following chapter.

3 Interface for AI Communication

In the previous chapter we discussed various challenges that artificial intelligence techniques face with communication and transparency, but also of the existing solutions for them. These solutions have been used to communicate the program's reasoning for the user (Lapuschkin et al., 2019). But what would be the most natural way to relay this? This is not a trivial task, as AI algorithms tend to process a lot of data, which creates a lot of information that may or may not be useful. Showing all of it at once can confuse and overwhelm the user. From this angle, the challenge of AI communication is closely related to the challenge of data-heavy interaction design.

One currently popular method for data-heavy AI communication is to not show the data to the user, but to store it in some manner. This way, the developer of the system can access the data later on if the AI is malfunctioning somehow. In 2016 it was already proposed in Germany that this kind of a "black box" system should be mandatory in self-driving cars, so that a team of experts would be able to determine what happened if a crash or some other accident occurred (Wacket et al., 2016). While these kinds of systems are important tools for maintaining public trust and understanding of AI technologies, they still run into the transparency issue that AIs have to an average user: The user does not necessarily gain any better understanding of the AI while using it.

3.1 Chatbot as a Tool for Communication

One approach to showing the information to the user is by utilizing chatbot functionalities for displaying new information in real time, as separate messages. Most people have an understanding about how chat user interfaces work because they are commonly used. The recent success of ChatGPT (Open AI, 2022), as well as some older examples such as Cleverbot (Carpenter, 2014), have shown that people in general are willing and capable of utilizing a chat user interface when using an AI technology. An example of ChatGPT chat interface can be seen in Figure 3.1. Most of the interface is reserved for the messages, with

the newest ones emerging from the bottom. Below that, there is a simple text input component, which allows the user to type in messages.

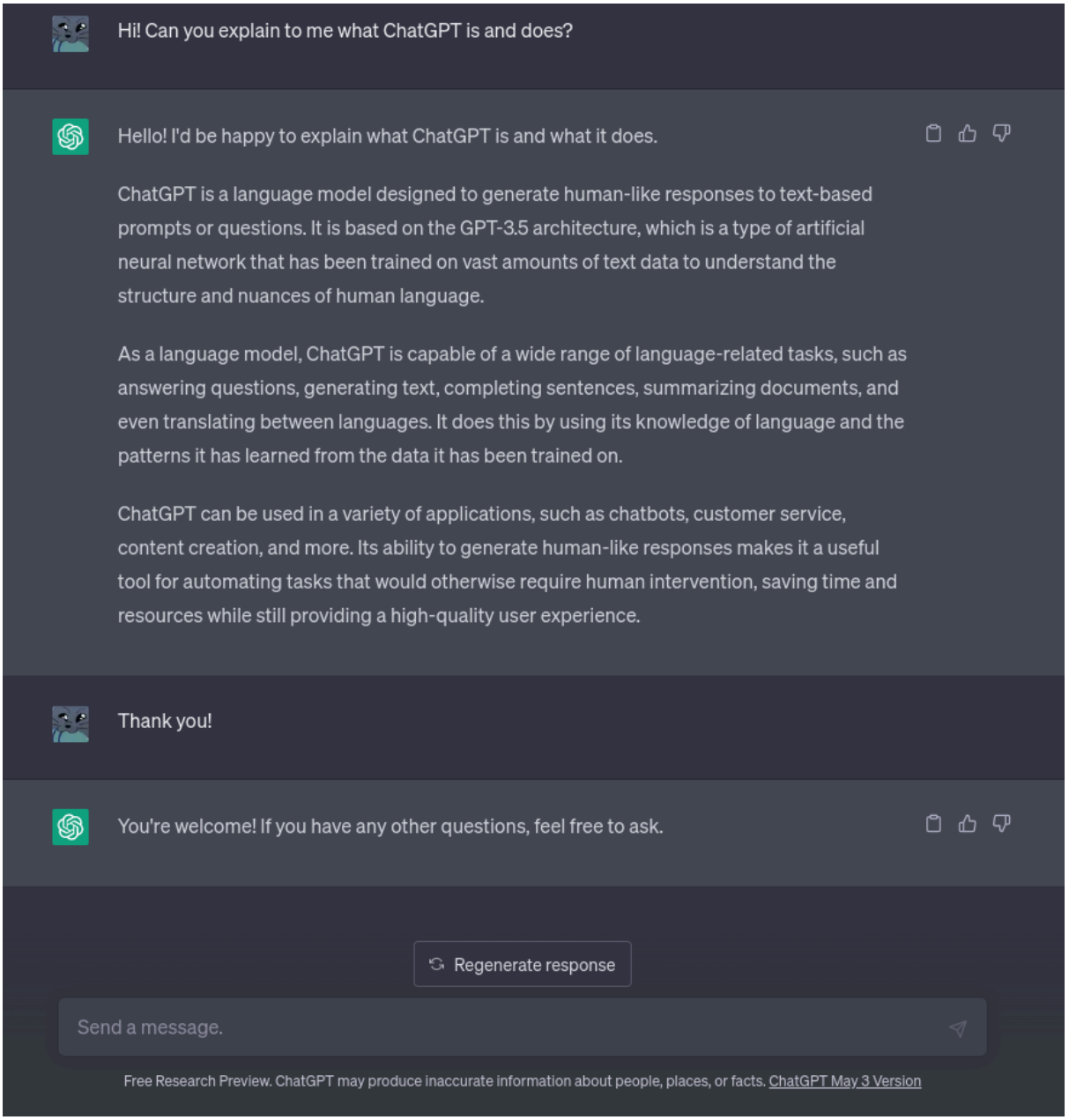


Figure 3.1: *A chat interface of ChatGPT in the free research preview. 5th of May, 2023.*

But in contrast to ChatGPT and Cleverbot, a chatbot using AI as a source of information should not be only reactive[4]. The information provided by the AI can be used to make the chatbot a proactive chatting partner that sends messages on its own, making it more

[4] A reactive chatbot refers to a chatbot that only responds to messages sent by the user. If the user does not send any messages, the chatbot does nothing. A good example of this is ChatGPT: In Figure 3.1 it can be seen that the chatbot only reacts to the user's messages, and does not write any more messages until the user has replied to the previous one.

independent from the user. Some research indicates that proactive chatbots have some advantages over purely reactive ones, due to the interaction not being dependent solely on the user (Almada et al., 2022). It is important, however, to pay attention to the amount of messages sent by the chatbot, so that it does not overwhelm the user.

As discussed before, a chat interface also provides the user a way to type in messages and send them to the chat. The chatbot could use natural language processing techniques to extract a meaning from these messages. This meaning could be, for example, a query for information after something has happened, a command that the AI recognizes, or just a non-functional chat with the chatbot. The processing can be very elaborate or very simple, but the user should nevertheless have a basic understanding of what agency they have over the AI when interacting with the chatbot. For example, if the user has the agency to stop the execution of the AI algorithm, the chat interface should communicate how to do it.

3.2 Planning a Chatbot Prototype

Using a chatbot to convey information about an AI algorithm prompted me to develop my own prototype for it, utilizing an open-source project called Mario AI Framework, created by Ahmed Khalifa in 2019[5]. The framework is a platform that supports playing AIs and level generation algorithms for the *Super Mario Bros.* videogame. Figure 3.2 shows the framework being run, with an AI playing the game. The videogame is essentially recreated for the platform, and is not emulating the original game. Because of this the platform is able to provide the AI developers an interface that is able to provide accurate information about the game state.

[5] https://amidos2006.github.io/Mario-AI-Framework/

Figure 3.2: *A screenshot of the Mario AI Framework running Super Mario Bros.*

Super Mario Bros. is a classic videogame used for many kinds of research in computer sciences. It is a platformer where you control a single character, called Mario, and navigate your way from the left side of the stage to the right side while avoiding obstacles and collecting coins and power-ups. Figure 3.2 contains the view of the game in the framework, with Mario in the middle of the screen. The game has been proven to be NP-hard (Demaine et al., 2016), which means that it can be used as a testing ground for complicated optimization algorithms. At least partly for this reason the game has been popular for AI developers as well, which means that there are existing AI algorithms to use for it. It also made the development of the chat interface and chatbot more simple, because I did not have to make the AIs for it as well.

The prototype that I made is called Chat Interface to Mario AI Framework, and its documentation can be found in Appendix A. The objectives for the prototype were the following:

1. Creating an interface that receives additional information for the decisions of an AI agent.
2. Creating a chat component on the other side of the interface, and saving the information in there.
3. Displaying the information in the chat without overwhelming the user, with the help of a chatbot.

Less effort was put into the chatbot functionality, although it was still important for the chatbot to feel like it was Mario chatting to the user. This is because chatbots also develop a sort of a personality, and making that personality Mario-themed felt fitting for the project.

3.3 Technical Review of the Prototype

On the user interface side, a chat component was attached to the side of the game view, as can be seen in Figure 3.3. The component contains the message log in the middle, three buttons below that and a text input component below those. The message log shows all sent messages between the user and the chatbot with timestamps, while the text input component is for sending messages. The three buttons switch the AI that controls Mario: The button with the "stop" symbol switches to a passive AI that does not move Mario, unless there is a danger nearby in which case it moves Mario to avoid it. The button with the "play" symbol switches to an AI that starts playing through the level cautiously, while the button with the "fast forward" symbol switches to an AI that tries to play through the level as fast as possible.

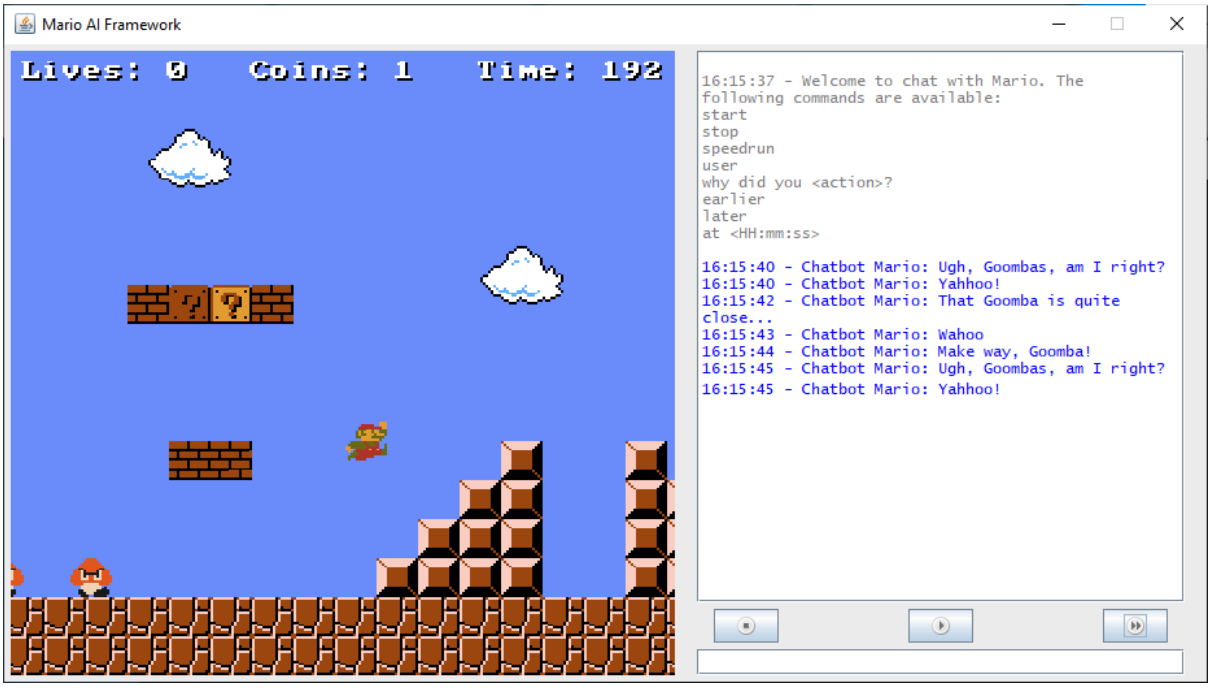


Figure 3.3: *A screenshot of Chat Interface to Mario AI Framework.*

Figure 3.4 depicts how the framework was expanded on: A chat system with a functioning chatbot was added to the framework that connects through an interface with the game, receiving the state of the game and additional information, called context, from the AI. This information is saved by the chatbot as a tree map[6] and sorted by their timestamp. The chatbot can access this tree map to find the reasoning of the AI for a control event that happened in that time slot, for example, if the AI made Mario jump. The AI does not need to provide context for everything; it is in fact completely voluntary for the AI algorithm to give any information. Of course the more information the AI is able to give, the more useful the chat interface is for the user.

[6] A tree map is an automatically sorted data structure, which guarantees a log(n) time cost performance for get, put and remove operations. See Java language documentation for the TreeMap class for more information.

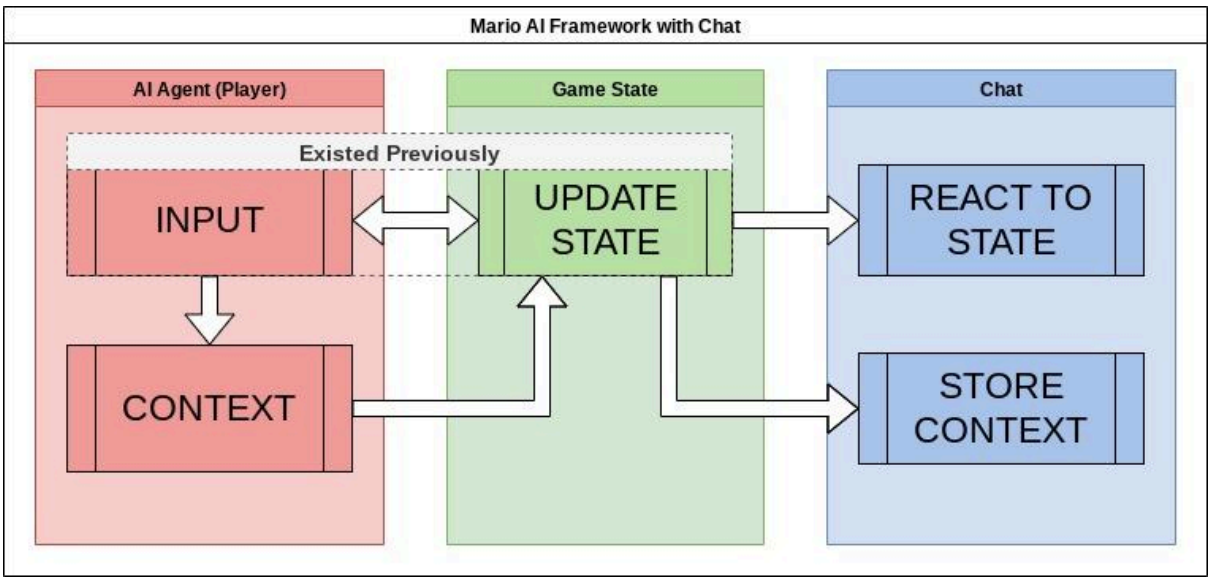


Figure 3.4: *A simplified diagram depicting the added functionality to the framework. The "Existed Previously" label refers to the functionality found in the Mario AI Framework by Ahmed Khalifa.*

3.3.1 Reaction Algorithm

Since it is possible that the AI does not provide information for the chatbot, the chatbot cannot be completely reliant on it. For this reason, I made a reaction algorithm to generate messages about the state of the game. This algorithm runs in real time while the game is playing, but it is also used to assess the given situation afterwards, by accessing the stored game state. This approach guarantees that the chatbot always has something to say, even if the AI did not provide proper information about its decision making process. The reaction algorithm essentially scans the current game screen for things of interest, and points them out. Special care was paid to detect dangers in front of Mario, such as enemy characters and holes in the ground. This algorithm is what makes the chatbot proactive, informing the user about things happening in the game.

The reaction algorithm in certain situations generated tons of messages. To address the issue of overwhelming the chat with them I created a bottleneck system to limit the amount of messages the chatbot is allowed to post. Every generated message is classified a type, and generating multiple messages of the same type in a short amount of time is blocked. For example, if an enemy is detected in front of Mario, it is announced only once until a fixed

amount of time has passed. This makes sure that the chatbot is still able to message about other relevant things, while keeping the number of messages at a reasonable amount.

3.3.2 User Interaction with the Chatbot

Additionally, I developed a query system to ask the chatbot for a reasoning to actions that the AI did during gameplay. For example, the user can ask in the chat "Why did you jump at 12:34:02?",[7] which causes the chatbot to search for jump events in the tree map around the specified time. If a time is not specified in the query, the chatbot fetches the last jump event that occurred. The user can also specify "earlier" or "later" if the chatbot does not find the right event, so that traversing the tree map is easier. After finding the event, the chatbot retrieves the reasoning of the AI attached to that event, and posts it if one exists. If a reason does not exist, the user is informed about this. After that, the chatbot also retrieves the game state for that event, runs the reaction algorithm with that state and posts the result. This can be useful when evaluating the AI's reasoning.

In order to make sure that the user would always be able to differentiate the reasoning given by the AI from the reaction algorithm of the chatbot, a coloring system was implemented that highlighted the reactions in blue and the AI reasoning in red, as can be seen in Figure 3.5. Messages sent by the user are highlighted in black color. In addition to this, sender names were added to the chat, with the messages generated by the reaction algorithm coming from the sender "Chatbot Mario" and the messages received from the AI coming from the sender "AI Mario". While this may seem excessive, it is extremely important that the user does not confuse the actual data coming from the AI and the chatter of the chatbot with each other.

[7] The time format HH:mm:ss was used because all the messages in the chat were timestamped using the same format. This means that the user can afterwards approximate when an event has happened based on the timestamps in the messages, and use them directly in the query.

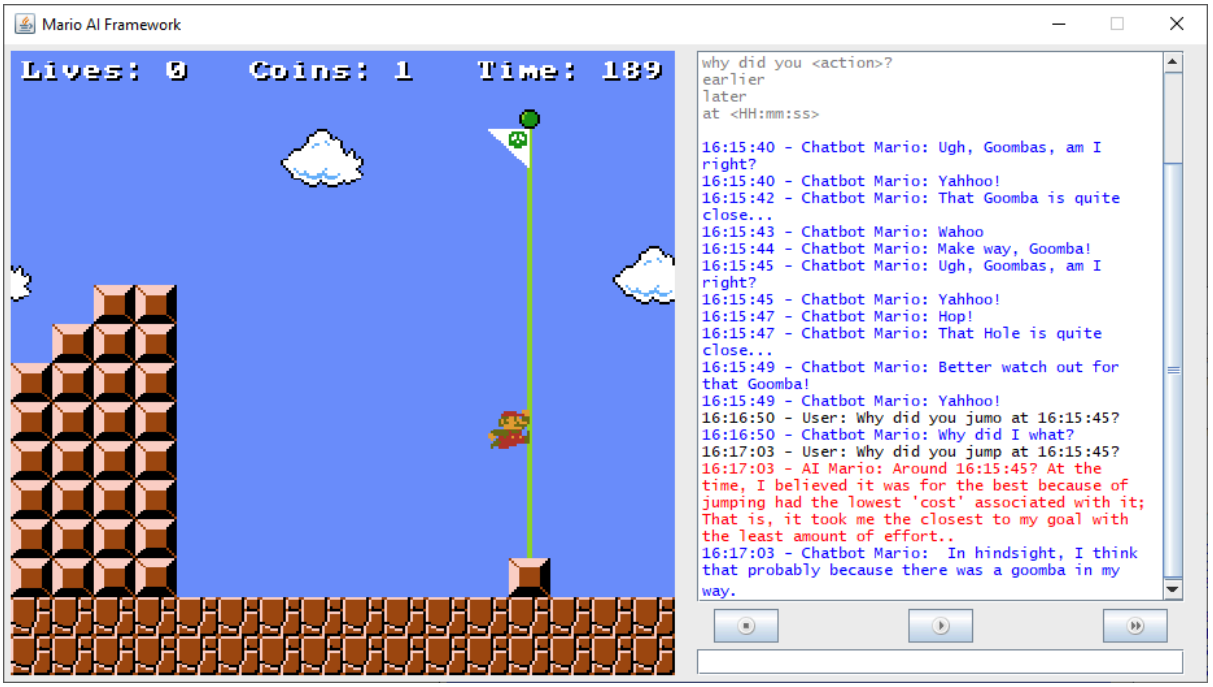


Figure 3.5: *The user asks a question from the chatbot.*

3.3.3 Limitations

One of the reasons for choosing the Mario AI Framework was that I did not have to make the AIs for running the game as well, but that choice was also the project's biggest limitation. It turned out that getting useful information from the ready-made AIs was very difficult when connecting them with the chat interface. Ultimately vague and general explanations had to be used for them, as can be seen from the "AI Mario" response in Figure 3.5. An alternative approach would have been to rewrite the AIs, but this was too much work for the scope of the project. This limitation highlights the difficulty of adding transparency to the AIs as an afterthought. Ideally, every AI project concerned with showing the reasoning of the system needs to take it into account all the way from the design phase.

Despite its challenges, the prototype showed that it is capable of storing and showing large amounts of information from various AIs playing *Super Mario Bros.* But is this useful for someone observing the AI? My initial hypothesis was that showing this information to the user would help them make more educated guesses about the performance of the AI, and reduce the amount of mystery about their inner workings. This hypothesis would still need to be tested in some manner.

Having developed the Chat Interface to Mario AI Framework from the Mario AI Framework, I was in the fortunate situation that there were essentially two versions of the framework, one with the chat interface and one without. This meant that some form of comparison between the two programs was possible, as they were otherwise identical. In the end I decided to do the comparison with a survey, as other people's experiences and feelings on AI and AI evaluation interested me the most.

4 Survey for Determining the Chatbot's Effects

The survey to evaluate the created prototype was hosted on Google Forms platform, and the link to it was distributed via email in various university channels and posted on social media (Twitter, Discord). The link was also shared in person during the preliminary presentation of my thesis topic. The survey was opened on the 20th of January 2023, and it was closed on the 31st of March 2023. Various ways to showcase the prototype to the responders were considered, but ultimately it was settled to be shown in video format. The results of the survey were saved in a Google Sheets file without email addresses or any other data that would compromise the responders anonymity. This chapter will cover the details of the survey in depth.

The aim of the survey was that the responders would evaluate the AIs playing in the Mario AI Framework in some manner, so quantitative questions were extensively used. The survey has some qualitative elements as well, but for the most part they are in place to catch the possible mistakes and misunderstandings of the quantitative sections. This chapter will first cover how the prototype Chat Interface to Mario AI Framework was used in the survey. Then, the structure of the survey is explained. And lastly, it will be disclosed how the survey was distributed.

4.1 Using Mario AI Frameworks for the Survey

My original idea was to have the responder use and play with the chatbot and then answer the survey. This method, however, had serious issues scaling up: Either I would have had to be present for each survey entry, helping the user installing and using the program, or alternatively an installer for the program would have had to be provided in the survey itself, with detailed instructions on how to operate it. The first approach would have been dependent on my schedule and availability, and the second approach would have required more time and effort to answer the survey, since reading instructions and trying to make a program work can take a lot of extra time. With two different programs, Mario AI Framework and Chat Interface to Mario AI Framework, I decided that installing and operating both of them for each survey entry was unfeasible.

Due to these challenges I decided to instead attach videos to the survey. The responder would first see Mario AI Framework being used in a video, and would then be asked to evaluate the AI playing *Super Mario Bros.* in the framework. After that they would see a similar video, but this time Chat Interface to Mario AI Framework would be used with the chat functionality, with the same evaluation questions. The AI playing the game and all the game parameters would be identical in the video, with the only difference being the chat features.[8] This way one could examine how the chat features affect the evaluation.

Using videos in the survey removes agency from the responder when compared to the alternative of letting them use the program before filling the survey. However, the videos make the survey faster and easier to respond to, which would improve the amount of responders. It also changes the perspective of the research: instead of talking about user experience, it would be more appropriate to talk about *observer experience*. An observer's experiences and feelings when seeing an AI system in action can offer valuable insight, especially since people can be led to form their opinions about AI based on videos and advertisements on social media. For example, at the time of writing this thesis, Tesla is facing lawsuits and criminal investigation due to releasing an allegedly staged video in 2016 about their car's autopilot feature, which made the people purchasing the car expect more from it (Jin, 2023). The objective of the survey would thus be to offer insight into how a chatbot integrated into an AI system affects how an observer views the system.

A total of four videos were selected to be shown in the survey, two without the chat features and two with them. Two different player "AIs" were used in the videos, with both of them being showcased with and without the chat interface. These algorithms were both very simple, with the first one simply choosing random, human-like patterns to move Mario with, and the other one being a modified A* search algorithm[9] that moved Mario towards the end of the level as fast as possible. It is important to note that in the videos the semi-random algorithm loses the game, while the A* algorithm gets to the end of the level each time. These algorithms were selected because they were essentially too simple to be considered as

[8] It is important to note, however, that identical parameters do not implicate identical gameplay. This is due to pseudo-random elements of the agents playing the game.

[9] A* is an efficient graph traversal and path search algorithm. However, it requires a lot of space, as all generated nodes are stored in memory. For *Super Mario Bros.*, this algorithm calculates the shortest possible path towards the goal without Mario taking damage. This path is then turned into inputs that the algorithm precisely executes.

AI solutions on their own, but had the potential to appear as such. A hypothesis with this approach was that the evaluation of the AIs would be more negative when viewing videos with the chat features, as the observer would see the insufficient reasoning of the AI.

4.2 The Structure of the Survey

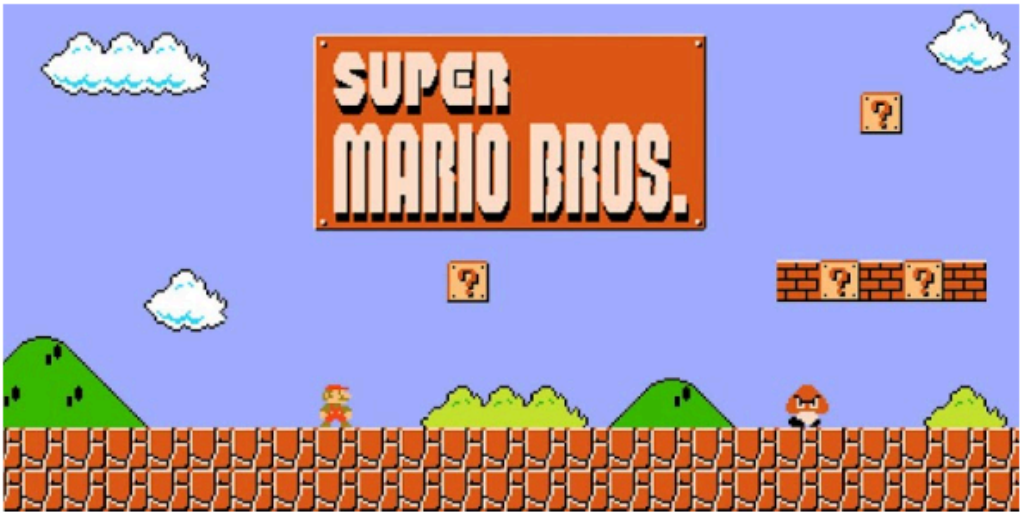
The survey itself consisted of three parts: general information, AI evaluations, and final questions. AI evaluations were repeated four times due to there being four AI videos to evaluate, making it the biggest section of the three. Nevertheless, it was estimated that completing the survey would take anywhere from seven to fifteen minutes, as the videos were short and the total number of questions was low. In the front page, the survey informed the responder of what the survey was about, the estimated time to complete the survey, and that the responder should be over 18 years old. The responders were also informed that the survey answers were anonymous and that no data collected by the survey could be used to determine the identity of the responder.[10]

4.2.1 General Information

General information section focused on getting data about the person answering the survey, such as their age and gender. For age, five options were given: From 18 to 24, 25 to 31, 32 to 40, 41 to 50, and over 50. These age groups were given to simplify the process of categorization by age when analyzing the answers. For gender, options were given for male, female, and other. Relevant questions related to the topic of the survey were also asked, such as familiarity with videogames and *Super Mario Bros.* Lastly, familiarity with artificial intelligence was asked, with the ability to select multiple options and a custom write-in field where the responder could fill in their own answer. An optional write-in field was given to elaborate on the familiarities. Having this kind of information about the responder helps with detecting if people with similar expertise on the relevant topics tend to answer the evaluation questions in the same manner. Figure 4.1 shows what some of the general questions looked like in the survey.

[10] This is of course, as long as the responder themselves does not write in the answers anything specific that makes them identifiable.

How familiar are you with the original Super Mario Bros. videogame? *



- ○ I have played the game extensively
- ◉ I have played the game a little bit
- ○ I have seen videos of someone else playing the game
- ○ I have seen/played other Mario games, but I have not seen or played Super Mario Bros.
- ○ I barely know anything about Mario

What is your familiarity with artificial intelligence (AI)? You may select more than one option. *

- ☑ I have done research involving AI
- ☑ I have used a program that utilizes AI techniques (e.g. ChatGPT)
- ☐ I have created or helped with creating a program that utilizes AI techniques
- ☑ I have read or watched about AI
- ☐ Muu: _____

Figure 4.1: *An excerpt of the general information section with example answers.*

4.2.2 AI Evaluations

The AI evaluations consisted of four identical parts, with one video present in the beginning of each part. The responder was asked to first watch the video, and then answer the questions below as seen in Figure 4.2. The responder was assured that they could rewatch the video as many times as they felt they needed to.

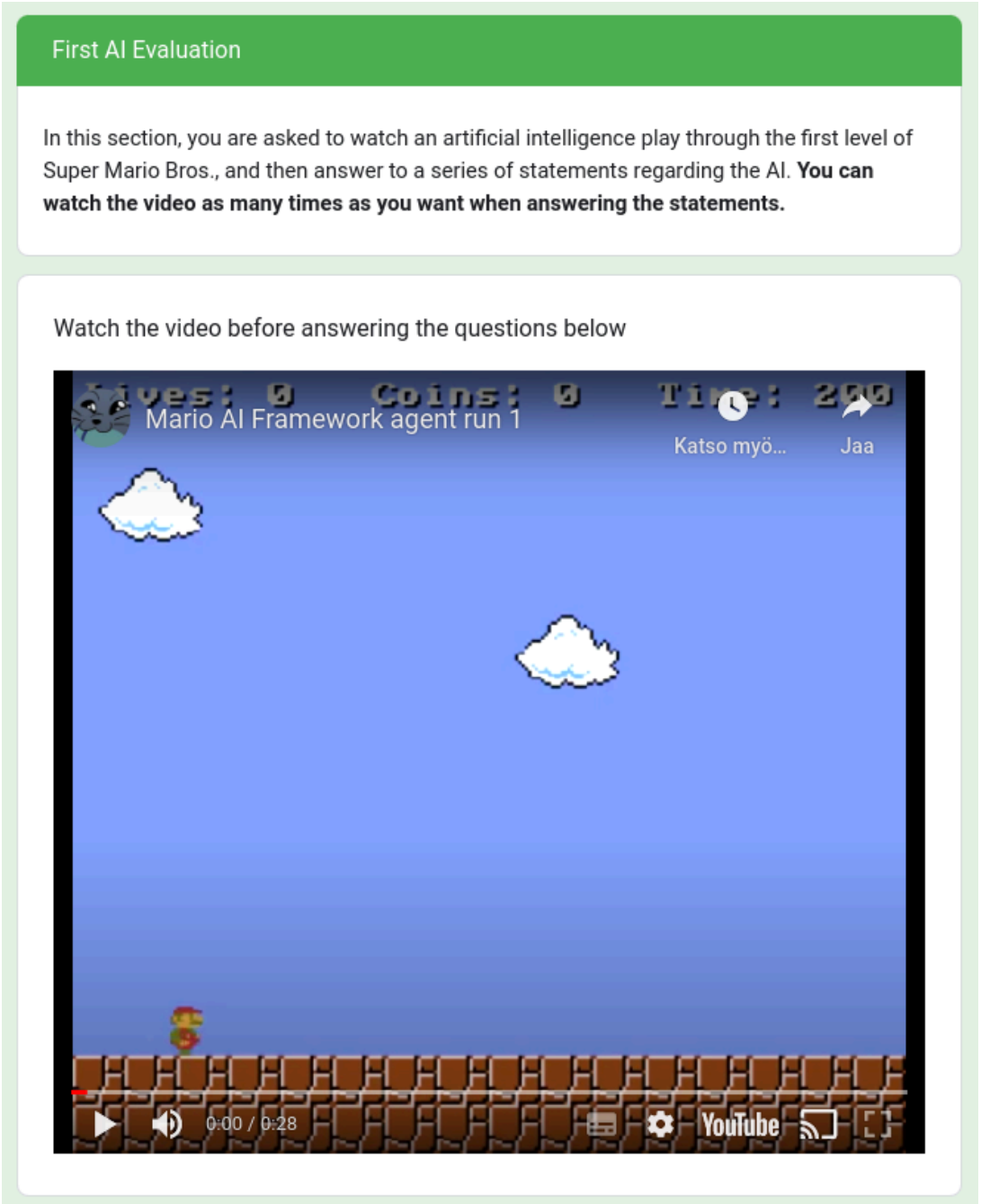


Figure 4.2: *The beginning of an AI Evaluation page in the survey.*

After the video, the responder was presented with eight statements regarding the AI that they could agree or disagree with on a Likert scale. This scale was numbered from 1 to 5, with 1 meaning that the responder strongly disagrees with the statement, and 5 meaning that they strongly agree. Number 3 represents a neutral stance, and the responder was also instructed to select this option if they were unsure (see Figure 4.3). The eight statements were the following:

1. *The AI achieved its goal*.
   This statement was a test to see whether the responder watched the video or not, or at least to see that the responder was understanding the game the same way that I was. For example, strongly agreeing with the statement when on the video the AI loses the game, would bring the validity of the responder's answers into question and cause concern if the survey was understood properly by them.

2. *The AI behaved in an intelligent manner*.
   This statement asks for the responder's impressions about the AI playing the game, and is the most direct statement in the survey for the responder to evaluate the AI's intelligence.

3. *The AI behaved in a human-like manner*.
   This statement's objective is to provide more variation into the question of intelligence. Despite the definition of intelligence being largely based on humans, it can still feel like something is intelligent but uncanny. Alternatively, it could reflect that something does not seem intelligent but still behaves in a human-like manner.

4. *The AI felt dumb*.
   This statement contradicts statement 2, meaning that the responder's evaluation of this statement should be more or less the opposite from statement 2. Its objective is to see whether the responder is consistent with their answers.

5. *I could play better than the AI*.
   This statement evaluates the performance of the AI in relation to the general information section, where the answer answered how experienced they were with

videogames and *Super Mario Bros.* This answer can be used to measure performance's relation to the perception of intelligence.

6. *The AI did something positively surprising*.
   This statement measures the optimism of the responder regarding the AI in the video. For example, the responder may have not considered the AI intelligent, but seeing silver linings indicates that the responder's attitude towards the AI is still positive.

7. *The AI did something negatively surprising*.
   In a similar way to the previous statement, this statement measures the pessimism of the responder regarding the AI in the video.

8. *The AI seemed to react to its environments*.
   The last statement tries to see if there exists a correlation between it and statement 2 or 3. If such exists, it could be speculated what reacting to one's environment is associated with more: intelligence or human-likeness.

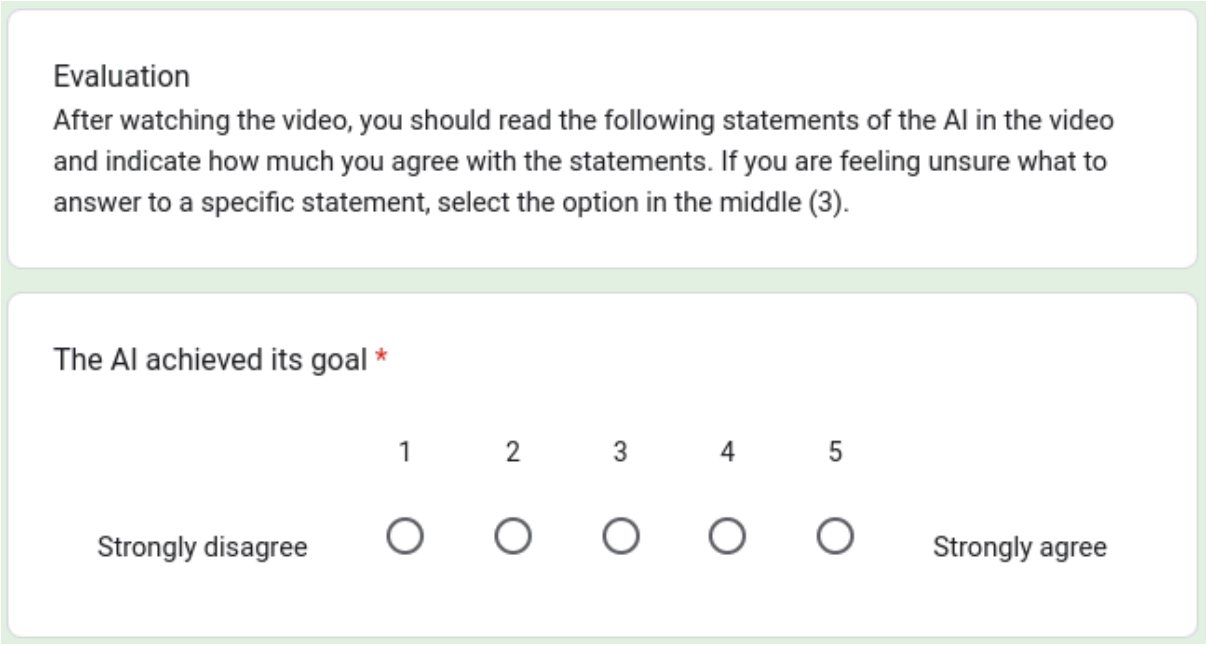


Figure 4.3: *The beginning of the evaluation section.*

After the evaluative statements, the responder was asked to reflect. They were asked two optional questions about how the evaluation felt like, and was there anything that felt unclear. Lastly, they were asked to rate how confident they felt when answering the questions and statements, as can be seen in Figure 4.4. The optional, open-ended questions were set to

figure out if there were issues or difficulties with the survey, and overall to acquire more qualitative data about the evaluation process. The confidence rating gives an idea about the overall feeling of the responder between the videos, and whether including the chat features to the video clips improved their confidence in evaluating the AIs.

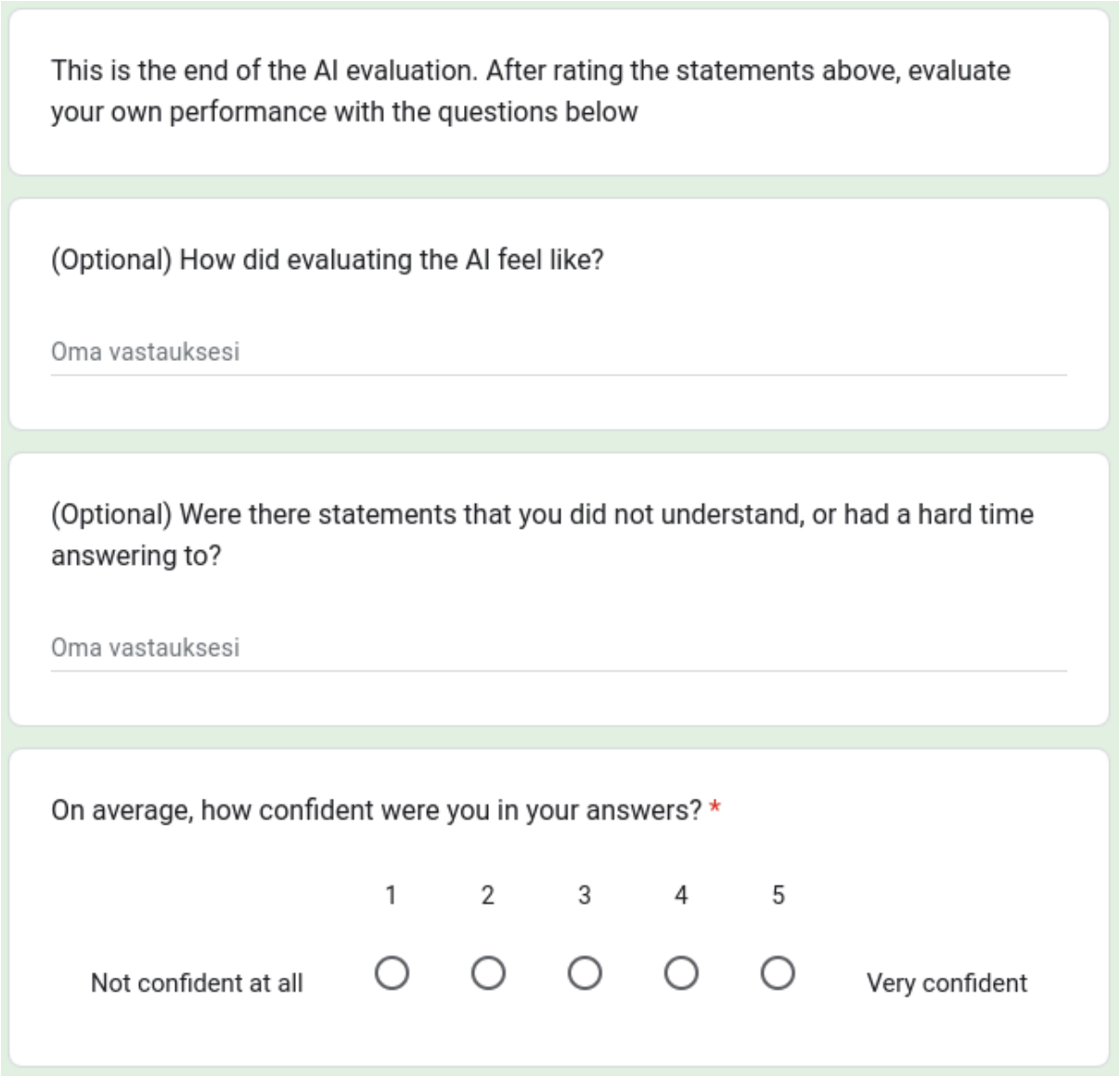


Figure 4.4: *Post-evaluation questions.*

The AI Evaluations section was repeated four times. On the third and fourth evaluations, the chat features were added to the videos. This was noted in the beginning of both evaluations, with instructions on how to read the chat. Additionally, a screenshot of the chat log was provided below the video for easier reading.

4.2.3 Final Questions

For the end of the survey, it was asked how many times on average the responder reviewed the video material when answering the evaluative questions. This question was at the end of the survey due to concerns that the responder would feel pressured to change their behavior if the question appeared in the AI evaluations section. The idea behind the question is to see if viewing the source material multiple times affects the evaluation. Lastly, the responder was given an option to give feedback about anything related to the survey. This is to gather information about possible oversights and things to improve.

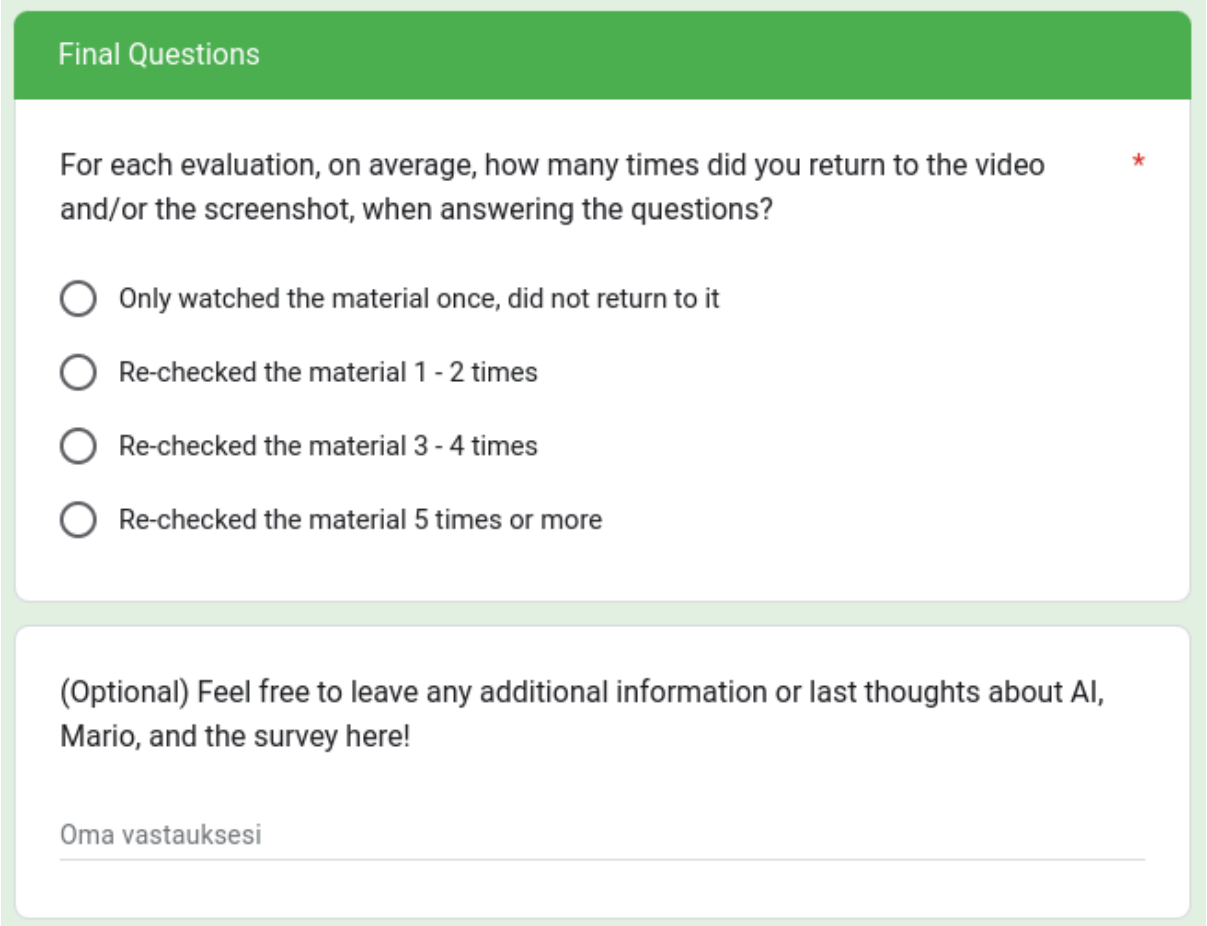


Figure 4.5: *The end of the survey.*

4.3 Distributing the Survey

The survey was hosted on Google Forms, an online form and survey creator. The platform allowed me to collect answers for the survey online, by sharing a public link to it. The link to

the survey provided by Google was long and full of random numbers and letters, however, so it was not ideal to show in university classes and in other places where one needed to type in the link to the browser. Because of this, a URL shortener service called Bitly was used to create an easy-to-read URL that redirected to the actual survey page.

Once I had the link for the survey, I wrote a small introduction to it and sent it to various places: to staff members of University of Turku who could share it on their courses or on their social media channels, to Twitter[11] on my own social media profile, and to multiple gaming-related servers on Discord with hundreds or thousands of members. The survey was in English, so it was not restricted to a Finnish-speaking audience. At the same time, international audiences tend to have more things going on, so it is easier for a single advertisement to get lost. I also asked the people in my interaction design seminar to answer the survey, during one of my thesis presentations.

Due to the distribution methods, the survey might have gotten more answers from students, and from younger people interested in videogames. As such, it may not be a representative sample of the general population. However, it still offers an interesting perspective from people interested in AI and videogames. Arranging a similar survey to a more neutral sampling would require a different approach, and a study of its own.

In the next chapter, the results of the survey will be analyzed. The objective is to gain an idea how the different videos affected the people answering the survey. For this end, the responders are categorized based on their general information, so that it can be seen if specific groups have tendencies to answer in a similar manner. It is examined if the chat features in the videos affect how intelligent the playing AI is perceived as. Finally, the self-evaluated “confidence rating” is analyzed, to see if the chat features increased or decreased the confidence of the responder.

[11] Recently known as X.

5 Analysis

The survey was conducted from 20th of January to 31st of March, 2023, and it received a total of 65 answers. The summary of the collected data can be found in Appendix B. This chapter will go through the collected answers, and determine whether the chat features had an effect on the survey responders evaluating the AIs. First, a criteria for the survey answers to be included in the survey is presented, and unsuitable answers are filtered out based on that. After that, demographic and other statistics of the survey responders are reviewed, which will be used to divide responders into groups. These groups and their responses are then compared with each other, in order to see if, for example, gender or prior knowledge in *Super Mario Bros.* have an effect on how the responders evaluate the AIs.

5.1 Filtering the Survey Responses

As discussed before, I decided to use the first evaluative statement in the survey, *The AI achieved its goal*, to gain an understanding whether the responders understood the performance of the AI in a similar manner. For me this question in the context of the survey would be similar to asking if Mario got to the end of the stage or not. My answers for this statement would have been 1 for the first evaluation, 5 for the second, 1 for the third and 5 for the fourth. I was not expecting every responder to answer in the same manner, but comparing these scores helped single out responders that might need to be considered to be filtered out from the analysis.

For most responders, their line of thinking seemed to be following my idea, as can be seen from Figure 5.1: The first evaluation featured a video of the AI managing through halfway of the stage and then running into a goomba[12], losing the game. It follows that the majority of responders, over 66 percent, have given a score of 1 or 2 on the Likert scale, as the AI did not beat the level.

[12] A goomba is an enemy character in the game. Mario takes damage if he touches a goomba, unless he jumps on its head.

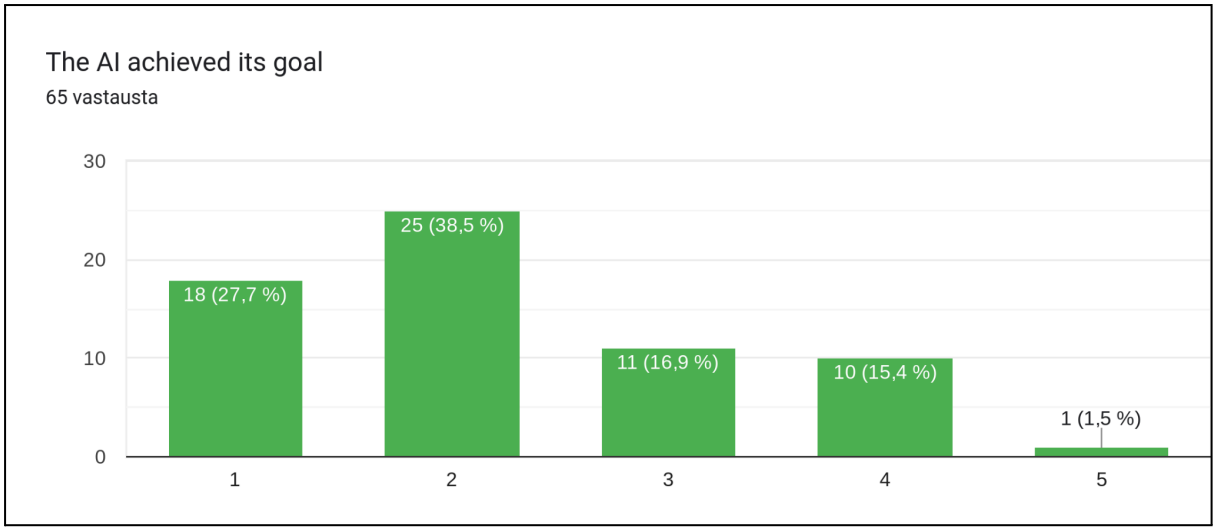


Figure 5.1: *Dispersion in the answers of the first evaluative section.*

On the other hand, surprisingly many gave the AI a score of 3 or 4, which indicates a neutral or slightly positive stance to the statement. This may be explained by this being the first evaluation in the survey. The AI got over halfway of the level, which might have made it seem like it did moderately well. This spread of answers was greatly reduced in the second evaluation, where the AI got convincingly to the end of the level, as can be seen in Figure 5.2.

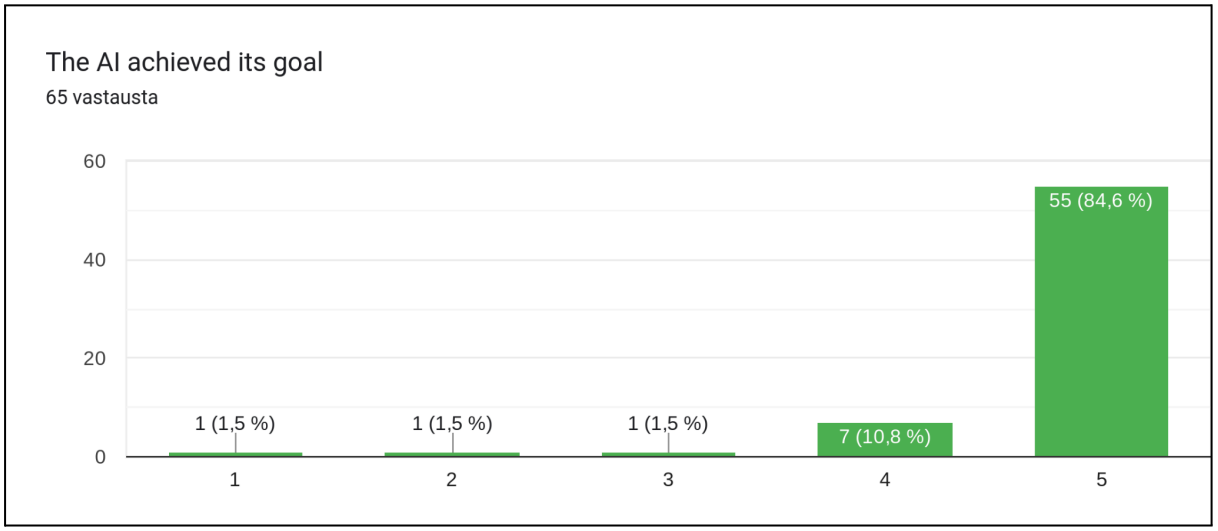


Figure 5.2: *Same question in the second evaluation.*

However, no matter which evaluation in question, responder #13 gave an exact opposite score from the majority for almost every single question. It is possible that they had either understood the Likert scale the wrong way around, or they were answering the survey wrong

on purpose. In either case, the answers of responder #13 are filtered out from the rest of the analysis. That leaves the amount of responses to 64.

5.2 General Statistics of the Survey

The questions regarding the demographic and the background of the responder were asked in order to see if the responders from one background answered differently to the survey when compared with others. In addition, it is also useful to know what kind of people the survey reached, and if some demographic groups were over- or underrepresented. Refer to Section 4.2.1 for an overview of all the questions analyzed in this section.

5.2.1 Age and Gender

From Figure 5.3 it is apparent that the survey attracted more responses from younger adults: The majority of the responders, 36 in total, were from 18 to 24 years old. The next biggest group, with 14 responders, was the second youngest group, from 25 to 31 years old. The ages 32-40 and 41-50 had 7 and 6 responders respectively, and there was only a single responder over 50 years old. The reason for this might be because the survey was for the large part advertised to university students and online videogame groups.

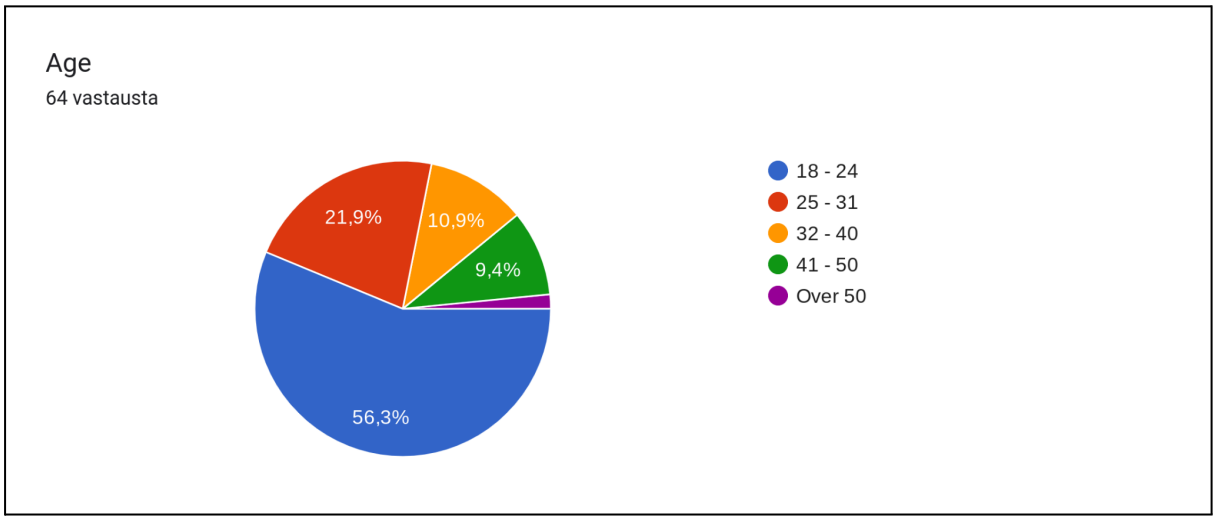


Figure 5.3: *Age variance in the survey.*

From the gender variance chart in Figure 5.4 we can see that the large majority of the responders, 40 in total, were male, while 15 of the responders were female and 9 in the other category. One idea on why men represent such a large portion of the responders is that in the

university circles, the advertisement for the survey probably circulated the most in the department of computer sciences. As an example, it can be seen from the website of University of Turku (Laaksonen, 2024) that out of the year 2022's 116 approved students to computer sciences, 77 were male and 39 were female. This 66-to-34 ratio of the faculty's gender distribution seems similar to the survey's, presuming that the gender ratio of attendants has stayed similar over the years. The university statistics do not permit other genders beyond male and female, so an accurate, direct comparison is difficult in this case though.

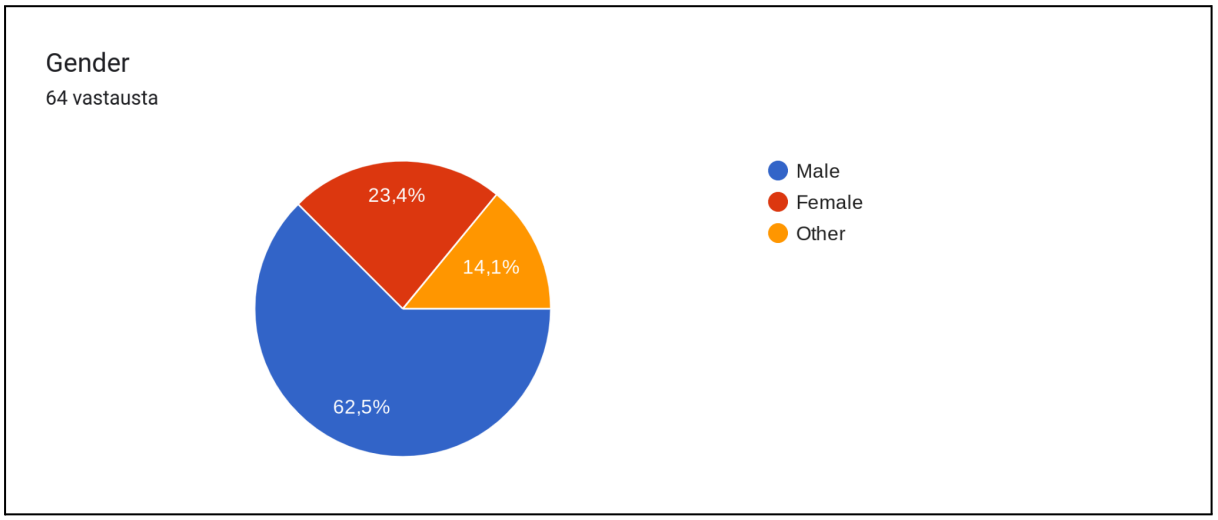


Figure 5.4: *Gender variance in the survey.*

When combining age and gender categories, it becomes apparent that men between the ages of 18 and 24 are overrepresented in the survey, being over one third of all the answers with 22 entries. Other combined gender and age categories are more evenly distributed. See Figure 5.5 for more details. This is important to keep in mind when considering the other demographic statistics collected by the survey.

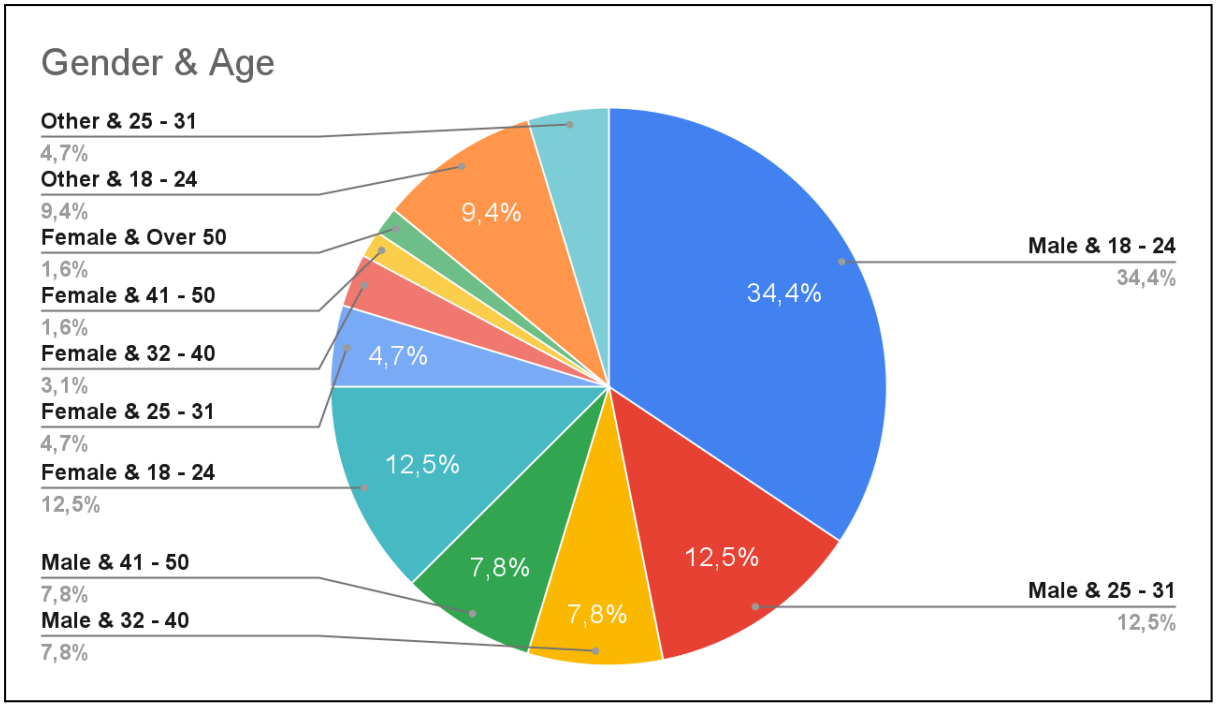


Figure 5.5: *Gender and age categories combined.*

5.2.2 Videogame Experience

Over 76 percent of the responders selected 4 or 5 when asked to estimate their videogame experience on a scale of 1 to 5, and only under 6 percent selected 1 or 2. This means that overall the survey reached people that were experienced in videogames. While this can introduce biases in the survey answers, it also has its uses: People experienced with videogames can be better at judging how an artificial intelligence is performing in a videogame, when compared to people unfamiliar with the medium.

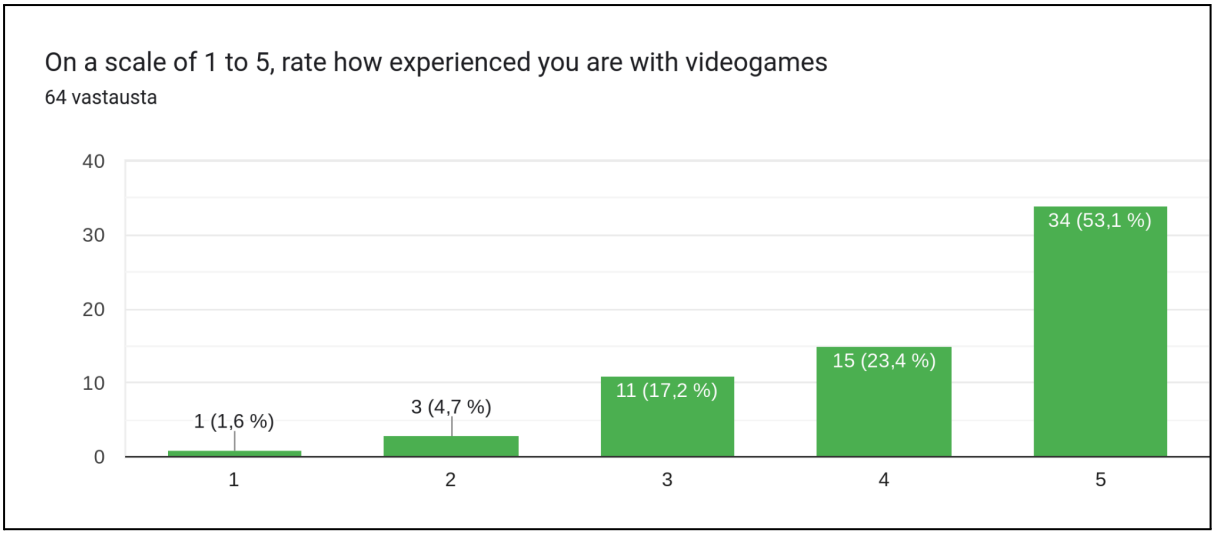


Figure 5.6: *Videogame experience of the responders.*

Over 90% of the responders had some experience with *Super Mario Bros.,* having either played or seen the game. Two of the responders reported that they were barely familiar with Mario, and one of them was not familiar with videogames or Mario. Cases like this are interesting as well, to see if unfamiliarity to these topics affects the evaluation. See sections 5.3.2 and 5.3.3 for the analysis of the differences. Figure 5.7 shows how familiarity with Mario was distributed.

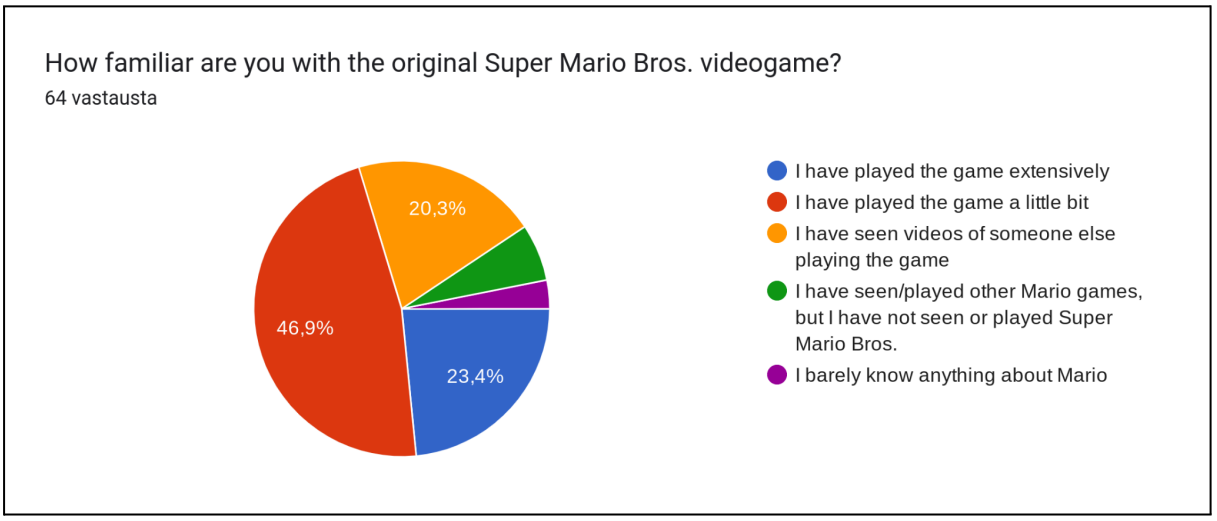


Figure 5.7: *Responders' familiarity with Super Mario Bros.*

5.2.3 Familiarity with Artificial Intelligence

The question about the responder's familiarity with artificial intelligence was different from the other statistical questions in two ways: First, it was a multi-choice question, meaning that

the responder could select multiple options instead of just one. Second, the responder could make a custom answer if they felt that none of the premade answers fit. Figure 5.8 shows the overall spread of the answers, as well as the two custom answers the responders have made.

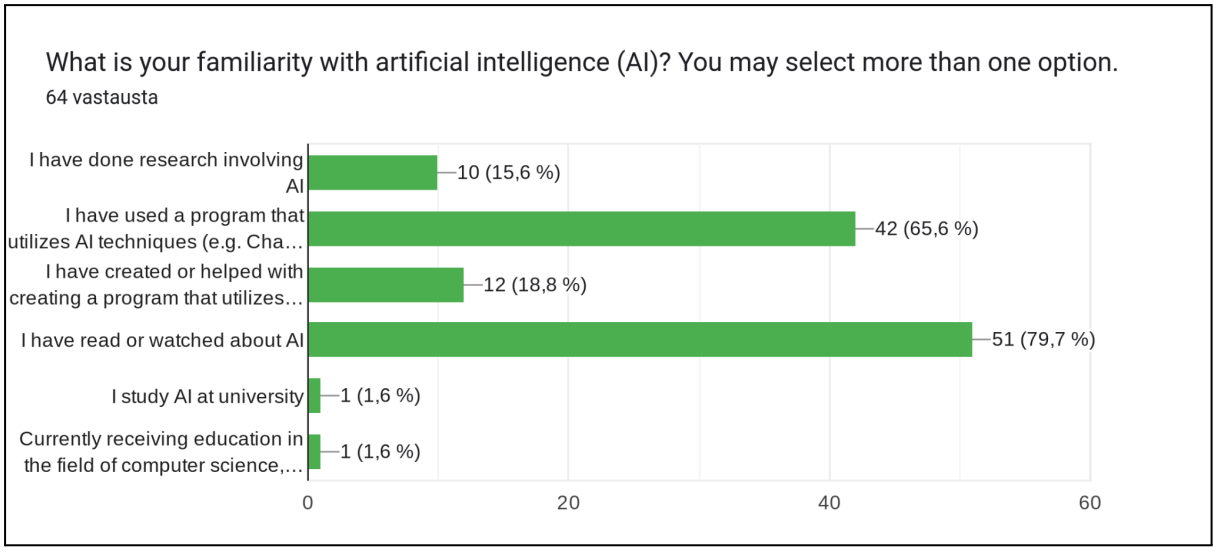


Figure 5.8: *Familiarity with artificial intelligence. The last two options are choices made by the responders.*

The custom answers made by the responders, *I study AI at university* and *Currently receiving education in the field of computer science, have passingly speculated on how to create an AI*, are both related to receiving education that relates to artificial intelligence. In hindsight, it would have been good to include an education-related option in the survey.

Most responders, almost 80 percent, answered that they had read or watched about AI. The majority of responders had also used an AI program, while researching and making AI programs was less common among the responders. No responders used the custom option, or the optional comment section, to report that they had no prior experience with artificial intelligence.

5.2.4 Number of Times Material Viewed

At the end of the survey it was asked that on average, how many times did the responder view the video material when evaluating an AI. These answers make the final general statistic that needs to be discussed. Figure 5.9 shows that most of the responders said that they re-checked the material once or twice, while the rest watched the material once for the most

part. Only four responders marked that they had re-checked the material three or four times. No responders reported that they had re-checked the material five times or more.

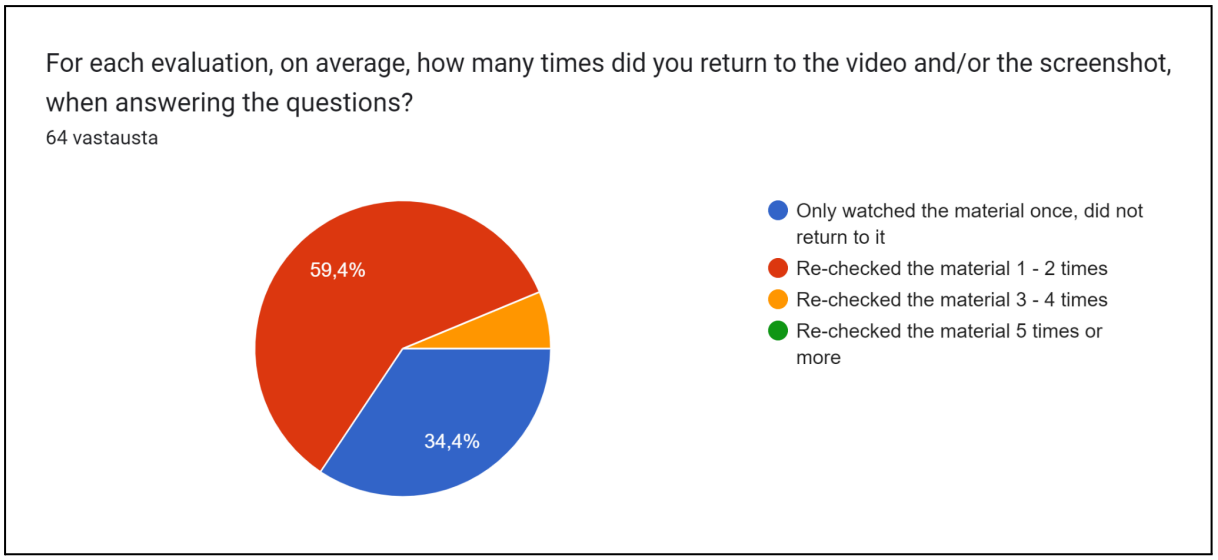


***Figure 5.9**: The vast majority of responders watched the material from one to three times.*

It seems that for the most part responders watched the material once or twice, but did not feel like they needed to review it that many times when answering the evaluation statements. This can be due to many things: Some may have been pressed for time, wanting to finish the survey as fast as possible. The responders might have also been generally confident in their answers, so they felt that they did not need to check the material again. This can be due to first impression bias, a psychological phenomenon where a person stops making new observations about the subject after a brief, initial exposure (Fang et al., 2020).

5.3 Analysis of the AI Evaluations

In this section the AI evaluation scores given by the responders are reviewed. The aim of this is to figure out how the inclusion of a chat interface affects the scoring, if it does. Additionally, by using the general statistics reviewed earlier, we can form groups of the responders and see how the groups' evaluations differ from each other. This may lead to insights on how things like gender, videogame or AI experience, or time used can affect the evaluation results. This section will first analyze all of the responses, and then move on to group analysis. For the purposes of comparing groups with each other, I decided that for a group of responders to stay comparable to other groups, it should represent at least five

percent of all the responders. In other words, a group needed to have at least four members. This was to avoid situations where a group was represented by a single responder. In some cases a group was joined together with another similar group, in order to create groups large enough for comparison.

Overall, the evaluation scores towards the two different AIs were within expectations. The first AI received unfavorable scores as it did not make it to the goal, while the second AI did much better in this regard. Considering that the first and third AI playing the game were the same in the survey, just like the second and the fourth were, the answers could be compared with each other to see if any scores were affected by the inclusion of the chat interface.

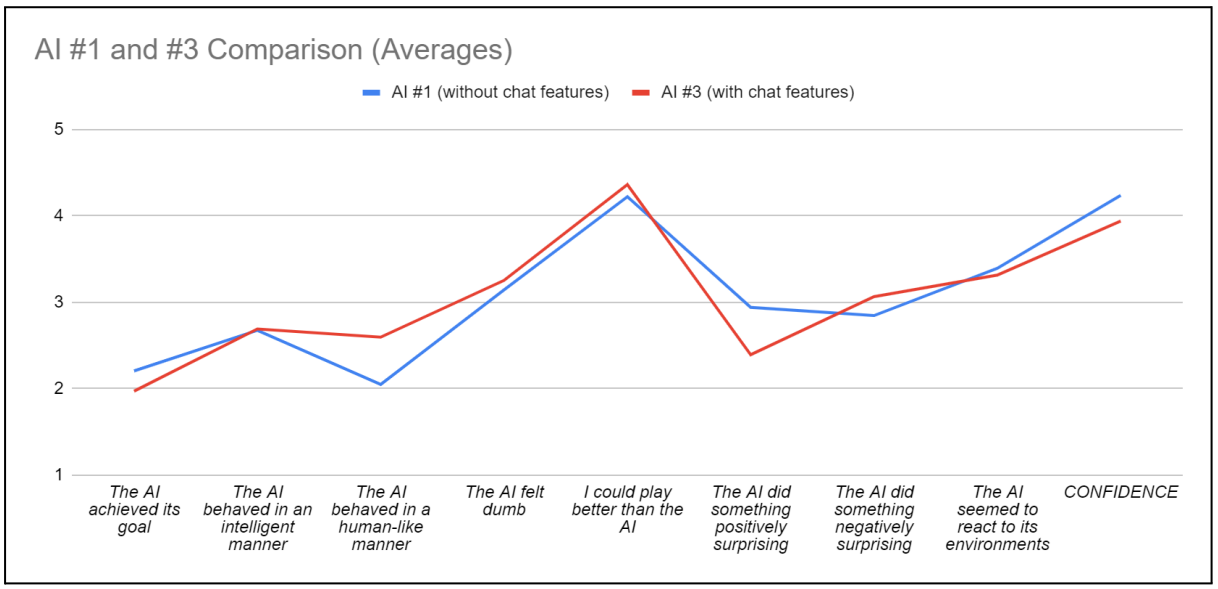


Figure 5.10: *Comparing the scoring of the first and the third AI*

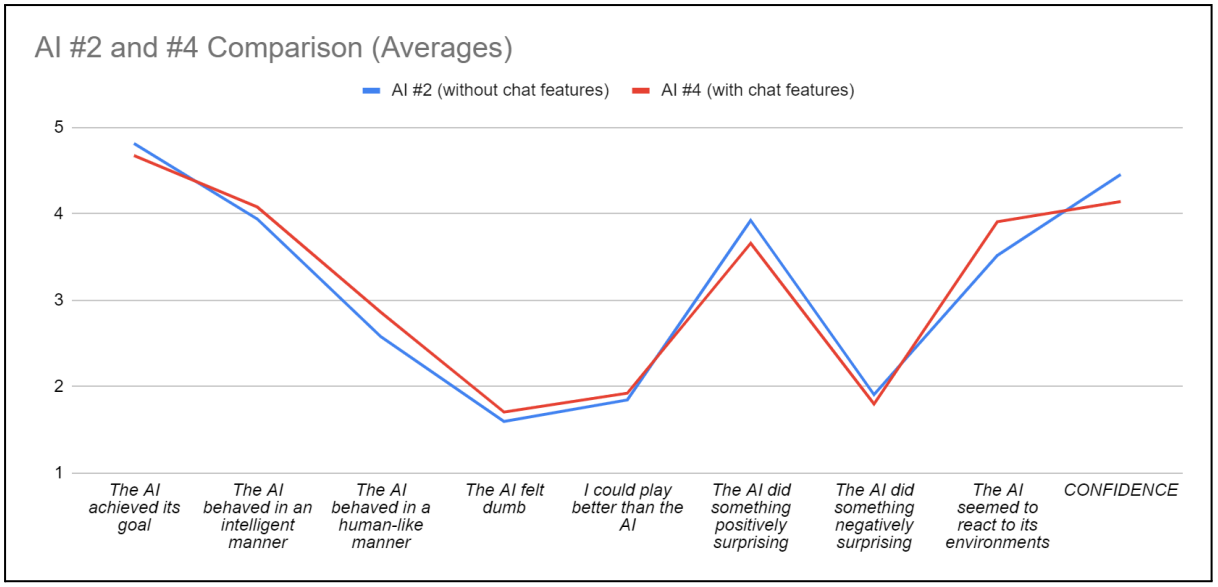


Figure 5.11: *Comparing the scoring of the second and the fourth AI*

There are several statements that seem to be affected the same way with Figures 5.10 and 5.11. The statement *The AI behaved in a human-like manner* is higher for both AIs when the chat features were in use, and the statement *The AI did something positively surprising* is conversely lower in both cases when the chat features were in use. This indicates that the chat features make the AI seem more human-like, while also taking some of the mystery away from how the AI operates[13]. In addition to these, the confidence score given at the end of the evaluation of the AIs was somewhat lower each time when the chat features were in use. This may imply that the chat features were confusing for some responders.

However, it is crucial to note that the differences between the AI evaluations were not very big. My initial expectation was that the responders' scoring of the AIs would turn more negative with the inclusion of the chat interface, as the interface communicates that the algorithms playing the game are not intelligent in the slightest. However, the data seems to suggest that this is not the case: The statement *The AI behaved in an intelligent manner* received roughly the same averages for the AIs with or without the chat features.

Comparing groups of responders based on the general statistics might shed more information on why the survey ended with these total averages. However, the group analysis produced a lot of data, a total of 24 charts from hundreds of data points, so showing it all in this chapter

[13] Although it is good to also keep in mind that seeing multiple AI evaluations in a row might affect a responder to be less positive about AIs and their performance in any case.

is not feasible. Because of this, the group analysis in the following subsections will only show a single graph per group comparison, which showcases the differences between the groups. The analysis itself takes all the data points into account. Refer to Appendix B to view all the charts generated for the group analysis.

5.3.1 Age and Gender

For differences between the gender groups of the survey, no considerable differences were found. For the most part, the men answering to the survey gave the highest confidence scores to themselves on average, although women had a higher average on the fourth AI evaluation. Combining the gender and age groups did not seem to affect the evaluations either, since no consistent differences between the gender and age groups could be found.

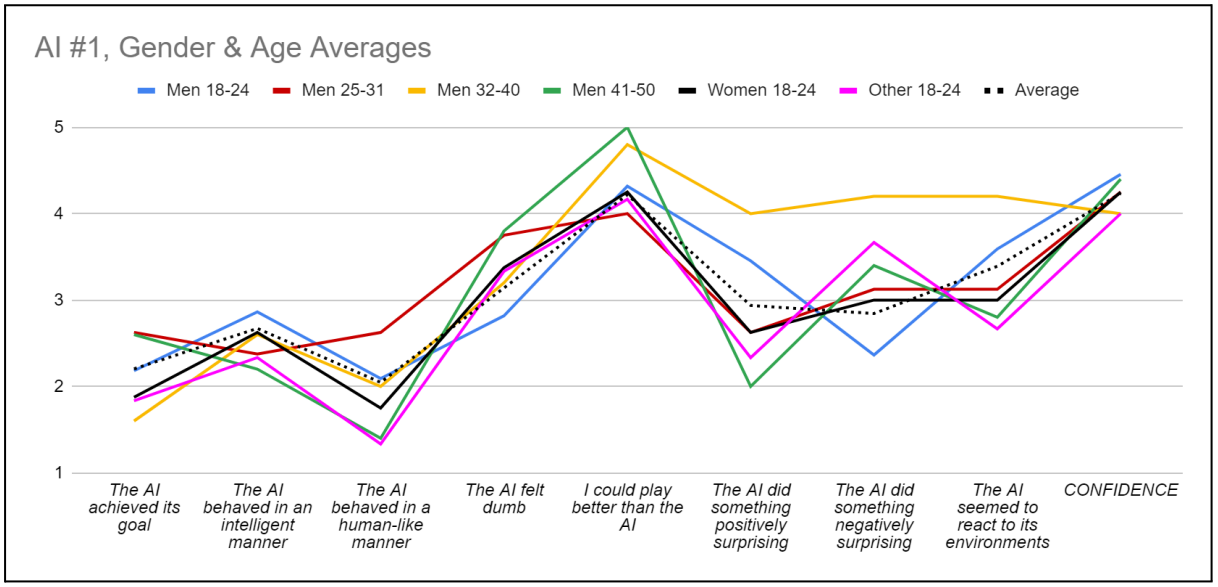


Figure 5.12: *Average scoring of the first AI by the responder's gender and age. Despite the differences in the averages of the groups, they were not consistent for the evaluations.*

5.3.2 Videogame Experience

For videogame experience groups, the responders that rated their experience to be 1 or 2 were joined into one group, so that they would create a large enough group for comparison. Curiously, this group with the least experience started with the lowest average confidence in their evaluations, but gained confidence very quickly: For the first evaluation, their average confidence amounted to 3.25, compared to the total average of 4.23, and by the fourth evaluation their average was 4.5, compared to the average of 4.14. This upwards trend in

confidence suggests that people unfamiliar with the task of the AI start out as unsure, but learn to trust their judgment very quickly when evaluating the AI.

This does not mean that they are more correct than others; In fact, the group's answers varied the most from the average consensus in this comparison. The group's variance from the total average was over 0.53, which is a 10.6% difference on the Likert scale. This is a significant average variance, for example, the group who had evaluated their videogame experience to be 3 had an average variance of 0.21. The prompt *The AI achieved its goal* stands out here, with an average variance of 0.77 for the group with the least videogame experience. This suggests that the group with very little videogame experience had a different idea about what the goal of the AI was, when compared to the other groups. See Figure 5.13 for an example of this, taken from an analysis of the fourth AI evaluation.

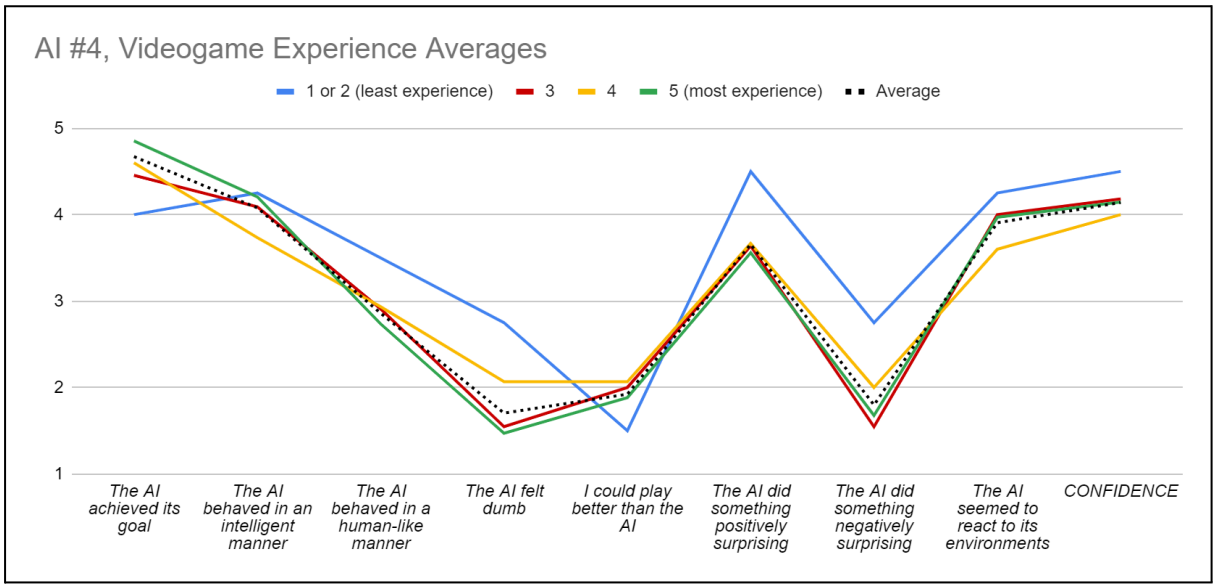


Figure 5.13: *Average scoring of the fourth AI by the responder's videogame experience*

5.3.3 Familiarity with Mario

In a similar manner to the videogame experience comparison, for *Super Mario Bros.* familiarity the groups *I barely know anything about Mario* and *I have seen/played other Mario games, but I have not seen or played Super Mario Bros.* had to be combined so that the group would be big enough for comparison with the other groups. This group would represent the people with the least experience with *Super Mario Bros.*

Prior knowledge with *Super Mario Bros.* seemed to have a more consistent effect on average scores than general videogame experience. The responders with the highest amount of experience were the most critical of the AIs in the first two evaluations, scoring lower with positive statements like *The AI behaved in an intelligent manner* and scoring higher with negative statements like *The AI felt dumb*. Responders with the least amount of experience had the lowest average confidence in the same evaluations, and their scoring varied the most from the average, in a similar manner to the group with least videogame experience. See Figure 5.14 to see the differences from the first evaluation.

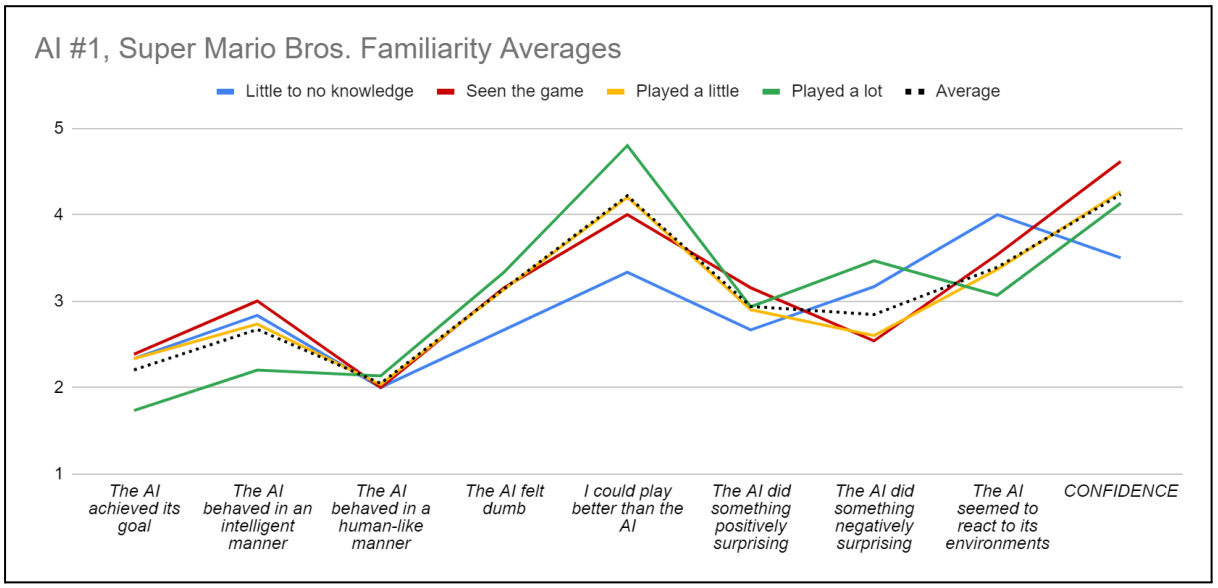


Figure 5.14: *Average scoring of the first AI by the responder's Mario experience*

However, this changed with the introduction of the chat features for the third and fourth evaluations. The confidence of the group with the least experience with *Super Mario Bros.* increased to be on par with the other groups, even though their answers were still the most varied. Also, the evaluation scores for the most experienced group became much more favorable towards the AI. For example, in the first evaluation the statement *The AI behaved in an intelligent manner* received an average score of 2.2 from the group with the most experience. For the third evaluation, with the same player AI but with the chat features added, the same group gave an average score of 2.67, which is a significant increase and much closer to the average score of all the responders. The group with the most experience also ended up being the least confident in their evaluations with the chat features.

It seems that the chat features and explanations actually made responders with prior knowledge and expertise less confident in the survey compared to other groups. This would explain the overall average confidence being lower in the third and fourth evaluation, since most of the responders were experienced with both AI and Mario. As discussed earlier, introducing new elements can cause confusion, even if the intention is to provide more information and help the user.

5.3.4 Familiarity with Artificial Intelligence

For the AI familiarity groups, the custom answers related to education, described previously in Section 5.2.3, were ignored. This was because only two people had made a custom answer, and both of them had also selected other options for the question. Therefore, every responder is included in the comparison, even if those two custom selections are ignored. Since this was a question where the responder could select multiple options, it meant that a single responder could represent multiple groups in this comparison.

In Figure 5.15 some of the differences between the groups can be seen. Namely, the responders who said that they had developed AI software had lower average confidence scores, and lower average scores for the statement *The AI behaved in a human-like manner*. This was contrasted by the group that researched AI, having generally very high confidence in their answers and a less pessimistic view on the AIs.

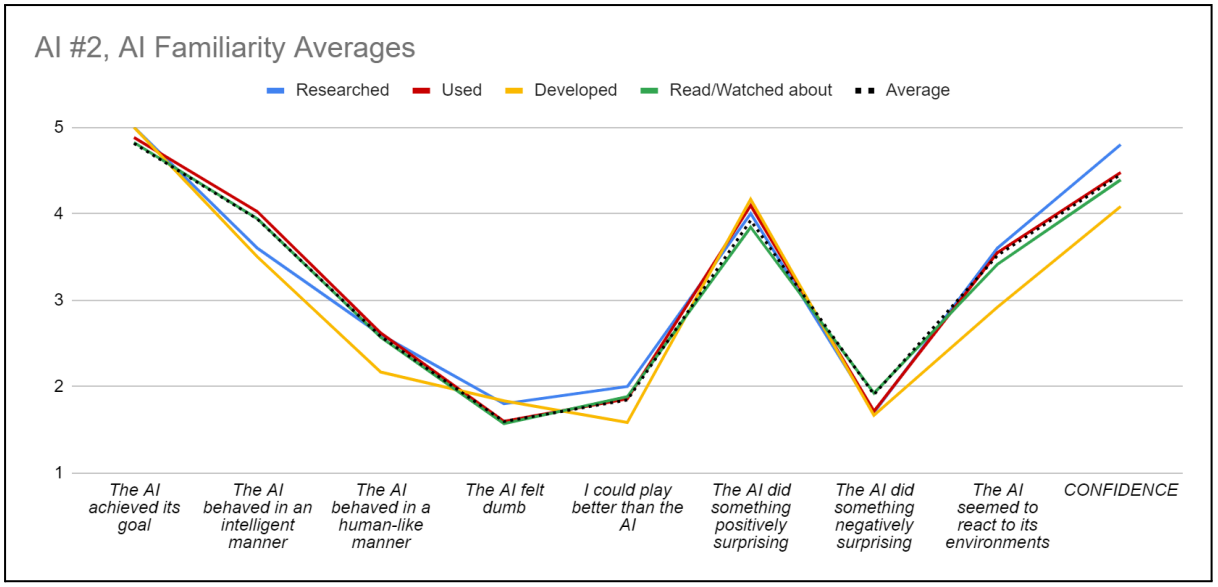


Figure 5.15: *Average scoring of the second AI by the responder's AI familiarity*

5.3.5 Number of Times Material Viewed

Lastly, it is considered whether the amount of times the video material was viewed affected the average evaluation scores. It is important to note that this data came from the responders' self-evaluation at the end of the survey, and as such was dependent on their memory and might not be accurate.

There was not much difference between the responders who watched the material only once and those who re-watched it once or twice. However, a change in the averages can be seen with responders who re-watched the material from three to four times: In all of the evaluations, this group had a higher average for the statement *The AI did something negatively surprising*, as well as the lowest confidence score. For example, see Figure 5.16. This can be assumed to imply that the responders who spent more time watching the AI often found something that they did not like, or understand about its behavior. As such, this factor seemed to affect the survey responders more than the inclusion of the chat interface. On the other hand, it could be argued that the group's confidence score was the lowest because they reviewed the material so many times. However, it does not disprove the correlation that the more times the responders reviewed the material, the more negative aspects they saw in the AI.

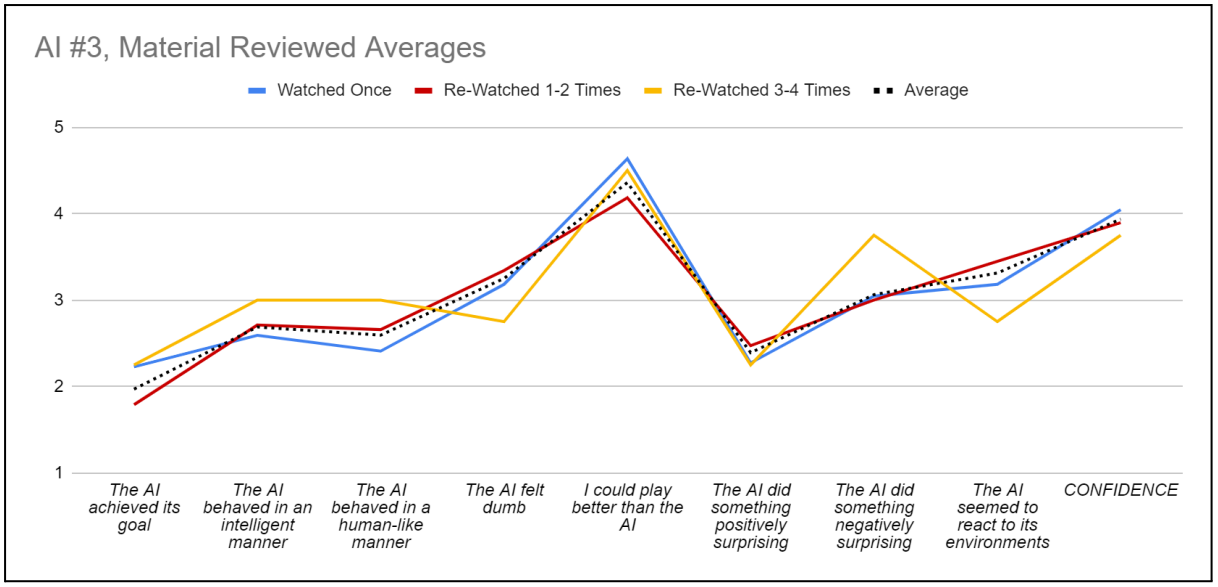


Figure 5.16: *Average scoring of the third AI by the times the material was reviewed*

5.3.6 Summary of the Analysis

The analysis of the data revealed small effects that the chat interface might have caused when evaluating the AIs. Comparing smaller groups with each other revealed more information about how differently various kinds of demographics see and evaluate the AI. The most important findings would be the following:

1. The introduction of the chat features seemed to lower the average confidence scores. Those familiar with the subject matter[14] lost their confidence the most with this change.
2. Responders unfamiliar with the subject matter start unconfident, but gain confidence quickly between the reviews, even surpassing the other groups. They can also have a different idea of what the goal of the AI is.
3. There was a correlation between the statement *The AI did something negatively surprising* and the number of times the AI material was viewed.

However, it needs to be reiterated that the effects of the chat features and the differences between the groups were smaller than hypothesized. A bigger study would be needed to

[14] That being videogames and *Super Mario Bros.* in this thesis.

confirm the correlations this thesis has found. There were also things that could be improved upon when conducting the study, which will be discussed in the next section.

5.4 Things to Improve

As mentioned before in Section 5.2.3, the familiarity with AI -question should have had an option for having studied artificial intelligence or a related field. In addition, the same question did not really benefit from letting the responders add custom answers. My original idea was that with enough custom options similar answers could be grouped up and studied, but in the end only two responders filled in a custom option. For future surveys, I would not allow custom answers for this kind of a question. The optional write-in fields were already used for the same effect by the responders to this survey, at least.

More analysis could have been done for the evaluation statements *The AI behaved in a human-like manner* and *The AI seemed to react to its environments*. In Section 4.2.2 it was discussed how these statements could be used to see how people see AIs as intelligent, and how much behaving human-like has to do with it. While the data was gathered about these statements, it was out of the scope of this thesis to do a further analysis. The data of all the responses is provided in Appendix B, and thus it is possible to return to this question in a later study.

Another thing the study would have benefitted from is more responses from people inexperienced with AI or Mario. With more inexperienced responders it would have been possible to make a more reliable analysis on how experience affected their evaluations. This could have been achieved by having more varied channels to spread the survey in, and also by making it even more clear that experience with AI or Mario was not expected to participate in the survey. One responder with little to no experience with Mario suggested in the survey: "*I think it would be nice to get a primer on the game Super Mario Bros at the beginning of the survey. I appreciate that I was probably not the target audience for this, but I think experience with Super Mario Bros might be recommended rather than just helpful.*" This leads me to believe that at least some of the inexperienced people felt unsure about taking part in the survey, which means that there may have been others who skipped on

answering the survey because of this. Having an optional introduction to the game might have alleviated this issue.

The findings in the analysis suggest small differences in reactions to AIs with different levels of explainability, but overall they were much smaller than expected. However, it can be hypothesized that what matters more is the duration of interaction or observation of the AI. This suggests that while transparency and communication can be important aspects of AI development in principle, the outlook of the results and the performance of the AI in the form of snippets and short videos still play a more convincing role for a typical observer. First impression bias that was discussed before may play a part in this as well (Fang et al., 2020).

6 Conclusion

The objective of this thesis was to highlight the difficulties of transparency with artificial intelligence. Some existing solutions to the challenges were explored in Chapter 2, but in Chapter 3 it was noted that these solutions are not really concerned with the delivery of information to the user. For studying this topic further, I developed the Chat Interface to Mario AI Framework to the existing Mario AI Framework (Khalifa, 2019), where I added a chat interface and chatbot functionalities to it. The chat interface felt like a natural choice of showing select information for the user, as well as allowing them to have some agency over the AI by issuing commands through the chat window.

To measure the effectiveness of the chat interface when it comes to users evaluating the competency of the AI, I created a survey with video material of both the Chat Interface to Mario AI Framework and the Mario AI Framework. By comparing the evaluation results of the two frameworks, it was possible to analyze the effectiveness of the chat interface on guiding users for more accurate evaluations of the performance of the AIs. The downside of using videos in the survey was that the responders were not able to interact with the AI or the chat themselves, which shifted their role from a user to an observer. However, this decision allowed collecting more answers for the survey, as a video format was much easier for the responder to go through than an executable file.

The survey gathered 65 answers before closing. One responder was filtered out from this pool, as their answers did not meet the control question criteria, and it was difficult to say whether they had watched the video material or not. From the remaining 64 responders, it was analyzed that the chat feature did not have a strong positive effect on the evaluation results. The chat features seemed to make the responders less confident in their answers, hinting that the responders had trouble understanding the information the chat was providing. However, responders who reviewed the video material three times or more seemed to evaluate the AIs more accurately.

6.1 Discussion

The development of artificial intelligence is currently in a boom, with its development seeing more commercial interest than ever before. Big companies like Amazon, Nvidia, Google and Microsoft are all investing in AI, as they are trying to claim the biggest piece of the market cake from the current AI technologies (CB Insights, 2024). This means that a lot of money is put into various companies trying to develop AI further, and those investments are expecting a return sooner rather than later.

It follows then that the current focus of AI development is to appeal to as many people as possible, in order to sell as many AI products and services as possible. The most common approach to reach the general public so far has been a cloud-based solution, where clients are able to utilize the AI via a web API or a website. The functionality of these approaches has been as generic as possible: Chat about anything with ChatGPT, form a picture about anything with Midjourney, or make a video about anything with Sora, and the list goes on. If an AI is at least a little bit useful for everyone, then the potential customer base includes everyone.

A cloud-based solution obfuscates how it functions, which can create temptations for the developers to cut corners: There are already examples of this happening, where actual people have posed as the AI chatbot (Contreras, 2024), or where some or all of the learning material of the AI was artificially mass-produced by thousands of underpaid workers (Haskins, 2024). This obfuscation is then used in marketing as well, where the focus is solely on the results of the solution and the convenience it brings. There can be a few technical words broadly explaining how the AI works, but they might not even be proved in any way.

Since AI solutions are often black boxes when it comes to explaining how they come up with their answers, it is natural to think that adding transparency to the process is a crucial next step for AI development. This is often called explainable AI, or xAI, and some companies have taken steps to this direction: For example, there is a research paper summarization tool called scite, that uses generative AI to create summarizations of a topic, but also shows citations from papers where it got the information from (Scite, 2021; Nicholson et al., 2021). The idea behind these kinds of solutions is that the user using the AI is able to verify the

information that the AI has given, therefore creating a more trustworthy and verifiable process. It would also help identify if there is something wrong with the AI.[15]

However, it seems that xAI is not effective in all cases. The analysis conducted in this thesis suggests that brief, controlled exposure to AI has a tendency of misleading the observer, even if it was not intended. Attempts at transparency with the chat interface providing information about the reasoning of the AI did not seem to help its observers much in making more accurate evaluations about it. One hypothesis is that the observer does not have the time to process all the available information in a short amount of time. In this context, things like xAI or legislation determining what information needs to be disclosed in fields such as AI marketing may not be effective.

On the other hand, the survey results analyzed in this thesis hinted that enough exposure to the same AI caused more uncertainty to the observers about its ability. This was a step in the right direction, as the AIs presented in this survey were fakes: The player AIs were just following simple, non-intelligent instructions. Therefore, longer overviews that showcase the AI more might be less prone to misleading the observer. Information overflow is a common effect in human psychology and can affect one's ability to evaluate things, but allocating more time to it alleviates the effect (Roetzel, 2019). It can be summarized that the more information is presented to the user, the more time they should be given to review it.

As AI development is embracing the commercial side, it would be important to consider these various ethical questions it brings. This thesis is not concerned with the ethical questions about AI itself, but rather with what happens when AI development and economic interests collide. When the development is the most concerned about meeting stakeholders' expectations, results become more important than the actual implementation. This can cause approaches where issues in performance are solved chiefly with more computing power. It also affects how the AI is showcased to the audience, as the results seem to be the only thing that matters; What, how much, and how fast.

Only focusing on the results leaves important pieces of information out. What data does the AI operate with? Does it store the information given by its users? And how much resources

[15] Interestingly, when looking for xAI solutions online on the 3rd of November, 2024, it seemed that there were not many products marketed using xAI. Furthermore, large companies were for the most part advertising that they have software solutions for only developing xAI.

does the building and upkeep of the AI require? It is unclear how much AI technologies increase the energy requirements of data centers (Desislavov et al., 2023), but it is clear that the need for these data centers, or computer farms, has increased with the popularization of large language models and other generative AIs (Hutchinson, 2024). And it is not just energy that these data centers need, but also space, water, and infrastructure. Microsoft's Azure AI may look like magic, but behind it lies large investments in space, infrastructure and technology – with the latest plans estimating the cost being around one hundred billion dollars (Bajwa et al., 2024). As a comparison, the current estimated costs for the company SpaceX's spaceship project, called Starship, are somewhere above five billion dollars (Sheetz & Kolodny, 2023).

In terms of scale it can thus be said that the largest current AI projects are comparable to outer space projects. However, what differentiates them from the huge AI projects done by Microsoft, Google and others, is that the outer space projects are monitored closely by the government that is affected by it, or in some cases by an intergovernmental organization like the European Space Agency.[16] Like AI, these projects take a lot of resources and contain a risk of causing global harm if something were to go wrong. It would therefore be reasonable to expect large-scale AI projects to be considered with a similar level of seriousness.

Artificial intelligence research has been a part of computer sciences for decades, and it has developed ideas into products and prototypes, like for example letter recognition for postal offices in the 1990s (Schank, 1991), speech recognition in the 2000s (Strait, 2023), and expert-level Go playing in the 2010s (Silver et al., 2016). But only in recent years has there been such a large effort to try and make AI products commercially successful for a wider audience. This brings new challenges and questions to the field, such as how to make sure nobody is cheating when competing to make the most money out of their product.

[16] As for why space programs face so much governmental control, see the Outer Space Treaty (United Nations, 1966).

All of this is curiously close to the problem with the Turing test, largely already discussed around the conception of the Loebner Prize competition[17], for example in the Special Interest Group on Artificial Intelligence Bulletin published by Association for Computing Machinery (ACM, 1992), when an algorithm was able to pass as a human for some judges by mimicking typing errors in its messages. It can be incredibly easy to fool people when they cannot see to the other side of the screen. In order to avoid this, AI software has to find a way to lift this veil, little by little, in the pursuit of solutions that exhibit actual intelligence. And the users who interact with the AI need to be given the time and means to process this extra information.

6.2 Future Work

The user experience side of artificial intelligence research is still in its infancy. A new focus is needed for researching and testing models suitable for showing data about the AI to the user, as this data is potentially numerous and contains the risk of confusing the user. Care needs to be put into filtering out duplicate information, while not hiding slight changes that might be of interest. One possible approach would be to have another AI take care of what data to show and what to omit, although this contains the risk of just moving the problem of obfuscation to another AI. No matter what the approach is, there is potential for interaction experience design and research to find novel models, so that in the future it would be easier to choose an interaction solution for specific xAI techniques.

There are also various aspects that were out of the scope of this thesis and would be beneficial to expand on. For one, as discussed in Section 5.4, the data collected for this thesis could offer more insights with further analysis. This is especially true when looking into statements regarding intelligence and behaviorism, such as *The AI behaved in a human-like manner* and *The AI seemed to react to its environments*. The full data set of the results of the survey can be found from Appendix B.

A similar user experience study could also be made for audiences not familiar with AI or computers. This would improve our understanding on what things affect different kinds of

[17] Loebner Prize competition was an event where computer programs were evaluated in a Turing test -like environment. The program that could convince the most judges of being a human typically won the competition, although the rules and the format of the competition changed regularly. The competition was held the first time in 1991, and it continued annually until 2019.

people when evaluating an AI. A qualitative approach might be more useful for a more varied group, to make sure that all the evaluators are on the same page.

Another avenue worth looking into in more detail is the first-impression bias with AI. It seems that the initial impressions with the technology leaves the observer with high confidence on average, but longer exposure times lower it. It would be interesting to confirm this with another study, with more details: What are the aspects that increase the confidence of the observer to believe they can make accurate judgements? How does exposure to a given AI's reasoning affect the confidence of evaluating it? Exploring this avenue more could give valuable insights into identifying so-called bad practices in AI marketing that have a tendency on misinforming or misleading the observer, for example.

Transparency of artificial intelligence most likely continues to provide challenges for years to come. From the viewpoint of this thesis it seems that these challenges come in multiple levels: Not only is it important to work on getting reliable information out of the decision making processes of AIs, but also bringing this information to the user needs consideration. Existing interaction design techniques can be used, but they also need to be evaluated again in the context of AIs. This aspect of AI research has not gotten much traction yet, and this thesis hints that it would be worth looking into, alongside other explainable AI research, in order to develop an AI that achieves good results while offering the user a new level of trust and verifiability.

References

Almada, A., Yu, Q., Patel, P., 2022. Proactive Chatbot Framework Based on the PS2CLH Model: An AI-Deep Learning Chatbot Assistant for Students. *Lecture Notes in Networks and Systems*, Volume 542. [online] Available at: <https://doi.org/10.1007/978-3-031-16072-1_54> [Accessed 3 November 2024].

Association for Computing Machinery (ACM), 1992. *Special Interest Group on Artificial Intelligence Bulletin*, Volume 3, Issue 4, pp. 7-11.

Bajwa, A., Simao, P., Gregorio, D., 2024. Microsoft, OpenAI plan $100 billion data-center project, media report says. *Reuters*. [online] Available at: <https://www.reuters.com/technology/microsoft-openai-planning-100-billion-data-center-project-information-reports-2024-03-29/> [Accessed 3 November 2024].

Basheer, I.A., Hajmeer, M., 2000. Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods*, Volume 43, Issue 1, pp. 3-31.

Bird, A.P., 2021. The Self-Driving Subway Train. *Medium*. [online] Available at: <https://alexand3r-bird.medium.com/the-self-driving-subway-train-b19eb80377f2> [Accessed 3 November 2024].

Braithwaite V.A., Ebbesson L.O., 2014. Pain and stress responses in farmed fish. *Revue Scientifique et Technique*, Volume 33, Issue 1, pp. 245-253.

Bridges, A.D., MaBouDi, H., Procenko, O., Lockwood, C., Mohammed, Y., Kowalewska, A., González, J., Woodgate, J., Chittka, L., 2023. Bumblebees acquire alternative puzzle-box solutions via social learning. *PLoS Biology*, Volume 21, Issue 3. [online] Available at: <https://doi.org/10.1371/journal.pbio.3002019> [Accessed 3 November 2024].

Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A., 2024. Video generation models as world simulators. *Open AI*. [online] Available at: <https://openai.com/research/video-generation-models-as-world-simulators> [Accessed 3 November 2024].

Carpenter, R., 2014. Turing Test: The bots are not amused. *Cleverbot*. [online] Available at: <https://www.cleverbot.com/amused> [Accessed 3 November 2024].

Carroll, M., Chan, A., Ashton, H., Krueger D., 2023. Characterizing Manipulation from AI Systems. *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. [online] Available at: <https://doi.org/10.1145/3617694.3623226> [Accessed 3 November 2024].

CB Insights, 2024. The big tech AI arms race: 75+ AI startups backed by Amazon, Google, Microsoft, and Nvidia. *CB Insights Research*. [online] Available at:

<https://www.cbinsights.com/research/report/big-tech-ai-investments/> [Accessed 3 November 2024].

Contreras, B., 2024. A Brief History of Automatons That Were Actually People. *Scientific American*. [online] Available at: <https://www.scientificamerican.com/article/is-there-a-human-hiding-behind-that-robot-or-ai/> [Accessed 3 November 2024].

De Cosmo, L., 2022. Google Engineer Claims AI Chatbot Is Sentient: Why That Matters. *Scientific American*. [online] Available at: <https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/> [Accessed 3 November 2024].

de Mouzon, C., Gonthier, M., Leboucher, G., 2023. Discrimination of cat-directed speech from human-directed speech in a population of indoor companion cats (Felis catus). *Animal Cognition*, Volume 26, Issue 2, pp. 611-619.

Demaine, E., Viglietta, G., Williams, A., 2016. Super Mario Bros. Is Harder/Easier than We Thought. *Proceedings of the 8th International Conference of Fun with Algorithms*, Article No. 13, pp. 1-15.

Desislavov, R., Martínez-Plumed, F., Hernández-Orallo, J., 2023. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, Volume 38. [online] Available at: <https://doi.org/10.1016/j.suscom.2023.100857> [Accessed 3 November 2024].

Fang, X., Rajkumar, T.M., Sena, M., Holsapple, C., 2020. National culture, online medium type, and first impression bias. *Journal of Organizational Computing and Electronic Commerce*, Volume 30, Issue 1, pp. 51-66.

Feingold, S., 2022. Artificial intelligence image generators bring delight - and concern. *World Economic Forum*. [online] Available at: <https://www.weforum.org/agenda/2022/10/ai-artist-systems-bring-delight-and-concern/> [Accessed 3 November 2024].

Gao, J., Tao, C., Jie, D., Lu, S., 2019. What is AI Software Testing? and Why. *IEEE International Conference on Service-Oriented System Engineering (SOSE)*, pp. 27-36.

Haskins, C., 2024. The Low-Paid Humans Behind AI's Smarts Ask Biden to Free Them From 'Modern Day Slavery'. *WIRED*. [online] Available at: <https://www.wired.com/story/low-paid-humans-ai-biden-modern-day-slavery/> [Accessed 3 November 2024].

Hernández-Orallo, J., Dowe, D.L., 2010. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174, pp. 1508-1539.

Huang, Y., Li, J., Fu, J., 2019. Review on Application of Artificial Intelligence in Civil Engineering. *Computer Modeling in Engineering & Sciences*, Volume 121, Issue 3, pp. 845-875.

Hutchinson, E., 2024. Growth of AI creates unprecedented demand for global data centres. *Intelligent Technologies*. [online] Available at: <https://www.intelligentdatacentres.com/2024/02/19/growth-of-ai-creates-unprecedented-demand-for-global-data-centres/> [Accessed 3 November 2024].

Jin, H., 2023. Tesla video promoting self-driving was staged, engineer testifies. *Reuters*. [online] Available at: <https://www.reuters.com/technology/tesla-video-promoting-self-driving-was-staged-engineer-testifies-2023-01-17/> [Accessed 3 November 2024].

Keneni, B.M., Kaur, D., Al Bataineh, A., Devabhaktuni, V.K., Javaid, A.Y., Zaientz, J.D., Marinier, R.P., 2019. Evolving Rule-Based Explainable Artificial Intelligence for Unmanned Aerial Vehicles. *IEEE Access*, Volume 7, pp. 17001-17016.

Kennedy, J., Eberhart, R.C., Shi, Y., 2001. Swarm Intelligence. *Morgan Kaufmann*.

Khalifa, A., 2019. Mario AI Framework | 10th Anniversary Edition. *GitHub*. [online] Available at: <https://amidos2006.github.io/Mario-AI-Framework/> [Accessed 3 November 2024].

Kim, Y., Sundar, S.S., 2012. Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, Volume 28, Issue 1, pp. 241-250.

Laaksonen, E., 2024. Hakijatilastot. *Turun yliopisto*. [online] Available at: <https://www.utu.fi/fi/opiskelutilastot/hakijatilastot> [Accessed 3 November 2024].

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, Volume 10, Issue 1096. [online] Available at: <https://doi.org/10.1038/s41467-019-08987-4> [Accessed 3 November 2024].

Legg, S., Hutter, M., 2007. Universal Intelligence: A Definition of Machine Intelligence. *Minds & Machines,* 17(4), pp. 391-444.

Lohr, S., 2016. IBM Is Counting on Its Bet on Watson, and Paying Big Money for It. *New York Times*. [online] Available at: <https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html> [Accessed 3 November 2024].

McNeill, M., Thro, E., 1994. Fuzzy Logic - A Practical Approach. *Elsevier Science & Technology*.

Newquist, H.P., 1994. The Brain Makers: Genius, Ego, And Greed in the Quest For Machines That Think. *Macmillan*.

Nicholson, J.M., Mordaunt, M., Lopez, P., Uppala, A., Rosati, D., Rodrigues, N.P., Grabitz, P., Rife, S.C., 2021. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, Volume 2, Issue 3, pp. 882-898.

Onyelowe, K., Mojtahedi, F., Ebid, A., Rezaei, A., Osinubi, K., Eberemu, A., Salahudeen, B., Gadzama, E., Rezazadeh, D., Jahangir, H., Yohanna, P., Onyia, M., Jalal, F., Iqbal, M., Chidozie Ikpa, Ifeyinwa I., Rehman, O., Rehman, Z., 2023. Selected AI optimization techniques and applications in geotechnical engineering. *Cogent Engineering*, Volume 10, Issue 1. [online] Available at: <https://doi.org/10.1080/23311916.2022.2153419> [Accessed 3 November 2024].

Open AI, 2022. Introducing ChatGPT. *Open AI*. [online] Available at: <https://openai.com/index/chatgpt/> [Accessed 3 November 2024].

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Roetzel, P., 2019. Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research*, Volume 12, pp. 479–522.

Samek, W., Müller, K.-R., 2019. Towards Explainable Artificial Intelligence. *Lecture Notes in Computer Science*, Volume 11700, pp. 5-22.

Saranya, A., Subhashini, R., 2023. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, Volume 7. [online] Available at: <https://doi.org/10.1016/j.dajour.2023.100230> [Accessed 3 November 2024].

Schank, R., 1991. Where's the AI? *AI Magazine*, Volume 12, Issue 4, pp. 38-49.

Scite, 2021. Scite Assistant - Your AI Research Partner. *Scite*. [online] Available at: <https://scite.ai/assistant> [Accessed 3 November 2024].

Sheetz, M., Kolodny, N., 2023. SpaceX set to join FAA to fight environmental lawsuit that could delay Starship work. *CNBC*. [online] Available at: <https://www.cnbc.com/2023/05/22/spacex-joining-faa-to-fight-environmental-lawsuit-over-starship.html> [Accessed 3 November 2024].

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, Volume 529, Issue 7587, pp. 484-489.

Stokel-Walker, C., 2024. ChatGPT is behaving weirdly (and you're probably reading too much into it). *Fast Company*. [online] Available at: <https://www.fastcompany.com/91033911/chatgpt-is-behaving-weirdly-and-youre-probably-reading-too-much-into-it> [Accessed 3 November 2024].

Strait, E., 2023. Understanding NLP History: The Evolution of Speech Recognition. *Lettria*. [online] Available at: <https://www.lettria.com/blogpost/understanding-nlp-history-the-evolution-of-speech-recognition> [Accessed 3 November 2024].

Turing, A., 1950. Computing Machinery And Intelligence. *Mind*, Volume LIX, Issue 236, pp. 433-460.

United Nations, 1966. Resolution 2222 (XXI): Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies. *United Nations Office for Outer Space Affairs*. [online] Available at: <https://www.unoosa.org/oosa/en/ourwork/spacelaw/treaties/outerspacetreaty.html> [Accessed 3 November 2024].

Ventura, S., Luna, J., Moyano, J., 2022. Genetic Algorithms. *Intech Open*. [online] Available at: <http://dx.doi.org/10.5772/intechopen.94664> [Accessed 3 November 2024].

Wacket, M., Copley, C., Mahlich, G., 2016. Germany to require 'black box' in autonomous cars. *Reuters*. [online] Available at: <https://www.reuters.com/article/us-germany-autos-idUSKCN0ZY1LT/> [Accessed 3 November 2024].

Wua, W.J., Lina, S.W., Moon, W., 2012. Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. *Computerized Medical Imaging and Graphics*, Volume 36, Issue 8, pp. 627–633.

Zhang, Q., Wallbridge, C., Jones, D., Morgan, P., 2024. Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents. *Transportation Research Part A: Policy and Practice*, Volume 179. [online] Available at: <https://doi.org/10.1016/j.tra.2023.103887> [Accessed 3 November 2024].

Appendices

Appendix A: Chat Interface to Mario AI Framework

Chat Interface to Mario AI Framework

by Villeveikko Sula (510918)

This document is an overview of my created chat interface for Ahmed Khalifa's Mario AI Framework. It explains the concept of the interface, as well as reflects on missing features and points of further development. The main methods that the chat uses are explained, as well as the interface between the AI agents and the chat. The aim of this document is to make the messages appearing to the chat as transparent as possible, show how the user is able to utilize the chat, and make future development easier.

Table of Contents

Overview of the Mario AI Framework

Mario AI Framework is a recreation of the *Super Mario Bros.* videogame, designed to make artificial intelligence development for the platform easier. It comes with ten agents that all play through the game differently, as well as a clear interface with which to add a new agent to the framework. The framework currently supports planning algorithms, with plans of supporting learning algorithms in the future. The framework also supports level generation algorithms, but covering those is outside the scope of this document.

Running the Framework

The framework is built with Java, so in order to run the game, you need to be able to compile and execute .java files. Using the command-line Java Development Kit (JDK) may be the easiest solution for this. To run the framework with the JDK, simply navigate to the "src" folder in your terminal, and type the following command:

```
javac PlayLevel.java
```

This will compile all files required by the framework. If no exceptions appear, type the following:

```
java PlayLevel.java
```

This should open a new window, in which you can see Mario trying to play through the first level of Super Mario Bros.

However, if you encounter an error while trying to run the game, it is probably because the framework is failing to read its image files. For Windows 10, I fixed the error in the following manner:

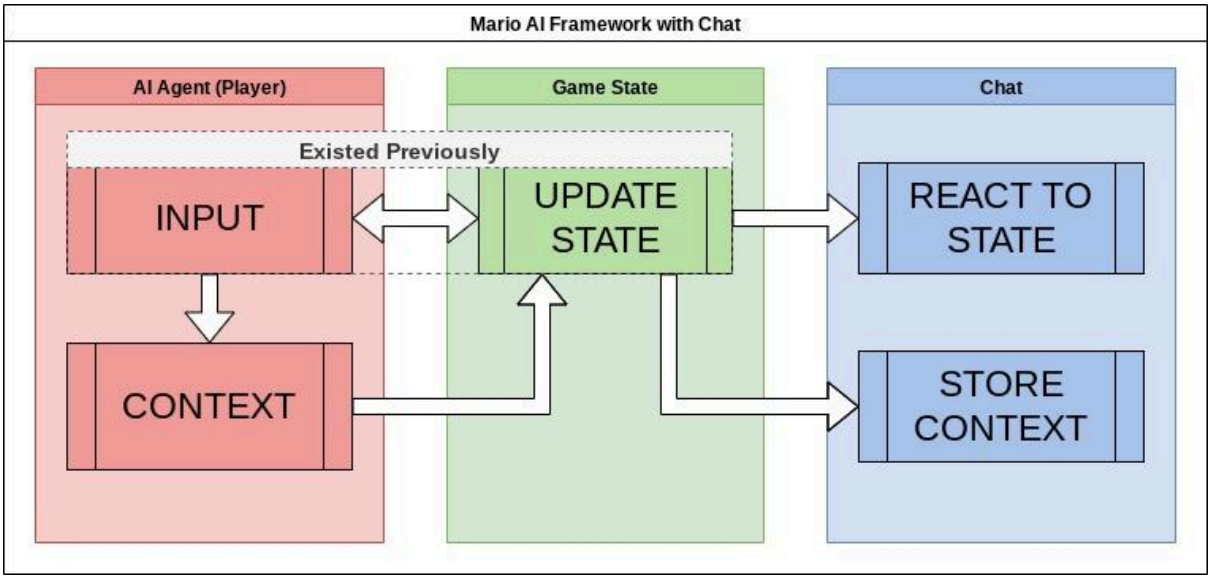
1. In file explorer, navigate to src\engine\helper folder and open Assets.java file in a text editor
2. Go to line 20, and edit `curDir + "/img/";` into `"../img/";`
3. Recompile Assets.java with the javac command. You may need to do this from the src

folder, in which case the command is `javac engine\helper\Assets.java` with your terminal navigated to the src folder of the framework.

If you encounter any other errors, it is probably because some file that the framework needs has not been compiled. Read through the error message, see what file(s) the message is referring to, and make sure that they are compiled.

To make modifications to how the framework runs, e.g. change the agent that plays through the game, open the file PlayLevel.java in a text editor and edit it. The readme contains instructions on which rows to edit and in what way.

Overview of the Chat Interface



A simplified diagram depicting the added functionality to the framework

The chat interface for the Mario AI Framework is a fork of the original repository, and works largely in the same manner. The intention behind this project is to add transparency to the agents playing through the game, as well as look into how communication affects the program's perceived intelligence. It has extended the functionality of the framework in the following ways:

- A chat log and an input window has been added to the user interface, on the right side of the game window. A chatbot that follows the current state of the game will react to its environments as “Mario”, while also responding to some of the messages that the user types in. Currently supported user prompts are:
    - Start/Go
    - Stop
    - Speedrun/Go fast
    - User/Take control
    - Why did you jump (at <HH:mm:ss>)?
    - Earlier/Later (than that)
- It is now possible to switch between the agents playing the game on the fly. To demonstrate this, there are three buttons in the chat interface that switch between three

different agents. It is also possible to switch to an agent by typing a message to the chat.

- Added a new agent to the framework, called “cautiousIdle”. This is essentially the same as the agent “doNothing” in the original framework, with the exception that the agent still tries to jump over the approaching enemies, and move to the right if it happens to be above a pit when it is activated. This agent makes stopping Mario from running in the middle of the level considerably safer.
- Added an additional function to the agent interface, that enables the agents to add a short message to each input the agent decides to do. This message is intended to give context or reasoning for the input, and it is stored with the timestamp of the input by the chat algorithm.
- In the chat interface, there’s a possibility to ask “Mario” why it jumped in a specific instance. In its answer, the chat utilizes both the message given by the agent playing the game, and the chatbot’s own interpretation of the game state at the specified moment. It does not yet support querying about other actions, but they should be easy to implement in the future. It can easily be distinguished which answer is from the agent playing the game, and what is just the interpretation of the chatbot.

Technical Review of the Program

The technical review is divided into two sections: In the first section, it is reviewed how the framework runs in general and its components briefly explained. This concerns largely about the original framework, but touches a little bit on how the chat interface is initialized. Level generation algorithms are not explained, as they are outside the scope of this project. The second section focuses on how the chat works and what interfaces have been built between the chat and the game. When referring to code and files in both sections of this review, this document refers to the chat interface repository, and not to the original framework's repository.

Mario AI Framework

We start examining the framework from the executable file PlayLevel.java and its main method at line 34. We can see that a MarioGame class is initialized and its runGame method is called (within a static printResults function). Let's review what this class contains in MarioGame.java, as it essentially handles the whole game.

There are two constructors to MarioGame, one where the constructor doesn't take any arguments (line 70) and one where it takes a parameter named killEvents (line 79). Currently only the empty constructor is used. Let's go over the static and instance variables of the class, as that will also explain the unused constructor:

- **maxTime** (static, long) - The maximum time that the agent playing the game is allowed to take for each input. In milliseconds.
- **graceTime** (static, long) - The "grace time" given by the program before it reports that the agent is taking more time than it should. In milliseconds.
- **width**/**height** (static, int) - The resolution of the game view.
- **tileWidth**/**tileHeight** (static, int) - the width and height of the game view in Mario game tiles. Calculated from the **width** and **height** variable.
- **verbose** (static, boolean) - Dictates whether the game should print out debug details to the console while the game is running.
- **pause** (instance, boolean) - Stops the game loop from running when set to true.
- **killEvents** (instance, MarioEvent[]) - Lists events (such as jumping, collecting a coin etc.) that will make the agent lose the game if they happen. This is normally left empty, but it can be used to give the agents additional restrictions. The constructor for MarioGame (line 79) accepts this kind of a list as a parameter, which then gets assigned to this variable.
- **newAgent** (instance, MarioAgent) - Contains a MarioAgent that the game should switch to on the next game loop. Otherwise null.
- **window** (instance, JFrame) - The game window that gets rendered when the game is run. Includes both the Super Mario Bros. game and the chat window.
- **render** (instance, MarioRender) - Handles the rendering of the Super Mario Bros. game world.

- **chat** (instance, MarioChat) - Handles the rendering and functionality of the chat instance.
- **agent** (instance, MarioAgent) - The current agent playing through the game.
- **world** (instance, MarioWorld) - Handles the state of the game.

With these variables, we have a basic understanding of what the class does. Next, let us review the function runGame that is called after instantiating the MarioGame class. We can quickly see that there are multiple overloaded runGame (and playGame) functions, which all end up to the following function signature:

```
public MarioResult runGame(MarioAgent agent, String level, int
timer, int marioState, boolean visuals, int fps, float scale)
```

The parameters of the function work as follows:

- **agent** is the AI that starts playing the game when the game initializes. This is set in the instance variable of the same name.
- **level** is the whole Mario level that is played through, described in text format. For more details, read the readme in the levels folder.
- **timer** sets how many seconds there are to play through the game.
- **marioState** sets the state of Mario for the beginning of the game. 0 equates to small Mario, 1 to large Mario, and 2 to fire flower Mario.
- **visuals** dictates whether the instance variable **window** will be initialized at all.
- **fps** sets the number of frames per second that the game loop function is following.
- **scale** is a multiplier for the actual screen size. For example, a value of 2 doubles the size of the window.

The function itself returns a MarioResult class, which contains the statistics of the current game. If the **visuals** parameter is true, the function will first render the game and the chat in the same window. It will then set the agent that is going to play the game, and then return the gameLoop-function:

```
private MarioResult gameLoop(String level, int timer, int
marioState, boolean visual, int fps)
```

This function was earlier referred to as the game loop. It uses a subset of the same parameters as runGame, so those do not need further explanation. In essence, the gameLoop function

initializes the instance variable **world**, renders the initial visuals with instance variable **render** if the local variable **visual** is set to true, and then enters a while loop. This loop repeats as long as the variable **world** states that the game is running.

In the while loop, the program will first get actions from the **agent** that is playing the game, which essentially amounts to which buttons the agent is pressing down in the game controls. Then, the game world will be updated according to these actions. The agent playing the game will be then asked to give additional context for its actions. After that, the chat will be updated with the events happening in the game world and with the context given by the agent. And lastly, the program checks if the **newAgent** variable has been updated, and switches to the given new agent if it has. If the instance variable **pause** is set to true, none of these things are done. But at the end, the loop renders the next frame of the game (if **visual** is set to true), and then sleeps depending on the **fps** set (if set to 0, the loop does not sleep at all, meaning that the game is executed as fast as your computer is able to process it).

Mario AI Chatbot

Now that we have a basic understanding on how the framework itself runs, we can take a closer look at how the chat and the chatbot are connected to it. In the runGame function that we discussed in the MarioGame class, it was mentioned that the chat is rendered with the game in the beginning of the function. This starts on line 218 with the MarioChat class being initialized, continues on line 222 with the chat being added to the JFrame window, and is finished on line 227 with the chat calling its init-method. Let's have a look at the MarioChat.java class:

Right from the beginning, we can see that MarioChat extends JComponent, which is the reason why it can be added to a JFrame window. Many of the instance variables of the class are components that make up the user interface of the chat, such as the text pane, the text field, the scroll pane and the three buttons. Exceptions to this are the following: **lastMessage**, which holds simple contextual information for the chatbot, **game**, which refers to the MarioGame class that the chat communicates with, and **chatWorker**, which refers to the ChatWorker class, working in a separate thread in the background as the chatbot of the system.

The constructor of MarioChat (line 54) is fairly simple, as it mainly just defines the properties of the UI, such as the dimensions according to the scale given. The init-method called later on (line 66) follows this up by initializing and drawing the chat graphics, adding the action listeners to the text input field and the buttons, and also adding the initial message in the chat log. Lastly, the method initializes and starts the MarioChatWorker thread.

MarioChat class takes care of pushing the messages in the chat view, as well as handling the user input in the chat pane in general. For example, if the user presses the "stop" button in the user interface, MarioChat will set cautiousIdle agent to MarioGame's newAgent variable (line 77). Handling the user's text input goes through the txtInputActionPerformed function (line 158), which will first add the user's message to the chat log, and then tries to parse a command from the user's message. The parsing happens in the parseUserMessageToCommand function (line 193).

The command parsing algorithm is currently very rudimentary, and would benefit from more development. Additionally, the chat does not react at all if it fails to parse a command from the user message. MarioChat could try to distinguish whether a given user input is a command or just friendly chatter, and then react accordingly.

In multiple instances when parsing commands from the user message related to queries about Mario's actions (lines 245, 255, 263), MarioChat calls methods from the MarioChatWorker class. Let's have a look at MarioChatWorker.java to gain more context.

MarioChatWorker is a thread class that receives the state of the game to the AddNewEventsToFunnel method, as well as the agent's reasoning for its inputs, each game cycle (as defined by the fps parameter in the gameLoop function in MarioGame). With this information, it both reacts to them in real time to the chat, as well as stores a lot of information (see **messageHistory** and **recentMessages**) so that it can be queried later on.

As a thread class, MarioChatWorker has a run method. This method is an endless loop that keeps emptying the local list **recentMessages** every 2000 milliseconds, or whatever value is defined to the local static variable **funnelRefreshInterval**. If recentMessages contains no items, the thread sleeps for 50 milliseconds (defined by the local static variable **funnelCheckupInterval**) before checking again. Essentially, the thread is used to keep cleaning up the recentMessages variable.

There are two functions in the AddNewEventsToFunnel method which create messages for the chat. The first one is TransformMarioEventsToMessages, which uses the received MarioEvent list to determine what has happened to Mario in the given frame, and generates messages out of those events accordingly. The second one is TransformForwardModelToObservations, which takes the MarioForwardModel that contains the state of the game and everything on the game view, and generates messages based on the things of interest in front of Mario: First it will check for holes, and then for enemies (goombas and koopas). Both of these functions can be improved quite easily by defining more things that generate chat messages.

After generating the messages in the helper functions, the AddNewEventsToFunnel method goes through each message. If a message of the same type exists in **recentMessages**, the

current message is not posted to the chat. If it does not exist, the message is posted, and it is then added to the recentMessages list. This structure ensures that the method does not spam the chat every frame of the game with similar messages.

Whether the message is posted to the chat or not, it is in any case added to **messageHistory**. As it is being saved, the **context** parameter is checked whether it contains additional data for the message. This is the data that an AI agent can give to justify itself for the actions that it is doing, for example jumping. If this data is found, it is added to the message. It should be considered whether this context-specific information should be formed into chat messages right away, instead of being just saved for later use.

There exists a function TransformMarioAgentEventToMessages, which takes the MarioAgentEvent list as a parameter, but does not yet create any chat messages. This is because MarioAgentEvent mainly contains positional information of Mario, and it is a bit unclear how this data could be used to generate messages. Utilizing this positional information in the future to create more accurate chat messages would be an interesting challenge.

MarioChatWorker also contains methods for asking for an explanation to a given type of event at a given timestamp. The default method for this is CheckHistoryForEventType, which takes the event type and the timestamp as parameters. Type must be given, and currently gives proper answers only to the “Jump” event type. Timestamp can be left as null, in which case the method will target the most recent event of the specified type. If the timestamp is specified, the method will search for a message of the given type in the local tree-based map **messageHistory** that happened the closest to the given timestamp, although the method will never search the events that happened prior to the given timestamp. For example, if the user wants to know why Mario jumped around 13:59:04, the method will never find an event that was recorded at the timestamp of 13:59:03.999, but will instead find an event that happened at 13:59:04.125, even though it is further away from the specified time. This is because the chat interface does not display milliseconds in its messages, so the user will almost always only search with timestamps rounded down to seconds, and it would be confusing to get an event of the previous second when searching for an event in the chat interface.

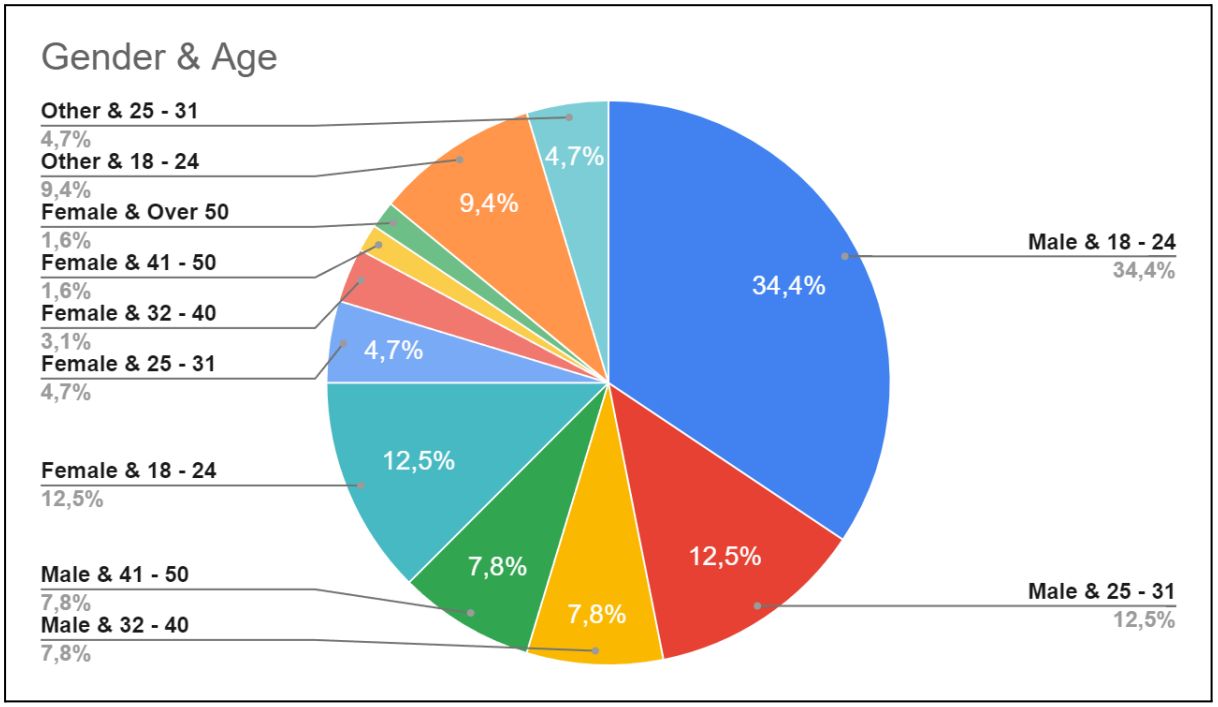
After finding a suitable recorded message from messageHistory, the function GiveIntrospectionForEvent is called to form the new chat messages. This function creates two messages: The first one contains the reasoning of the AI agent playing the game, if the agent has provided one at the time of the event. The second message calls for the CheckForDangersInFront function to check for reasons of the action afterwards, with the MarioForwardModel recorded within the message event. The CheckForDangersInFront function should probably be combined with the function TransformForwardModelToObservations, as they contain some duplicate code.

After the two messages are generated, they are returned to MarioChat to be posted to the chat log. When posting the messages, MarioChat colors the first message as red, and marks the message sender as “AI Mario”. All the other messages are colored blue, and the sender is marked as “Chatbot Mario”. The red messages are highlighted because their information comes directly from the AI agent, and as such, are more reliable. MarioChatWorker contains two other history functions, CheckEarlierHistoryForEventType and CheckLaterHistoryForEventType, which work like the described CheckHistoryForEventType, except that they are designed to search before or after the specified timestamp for matching events.

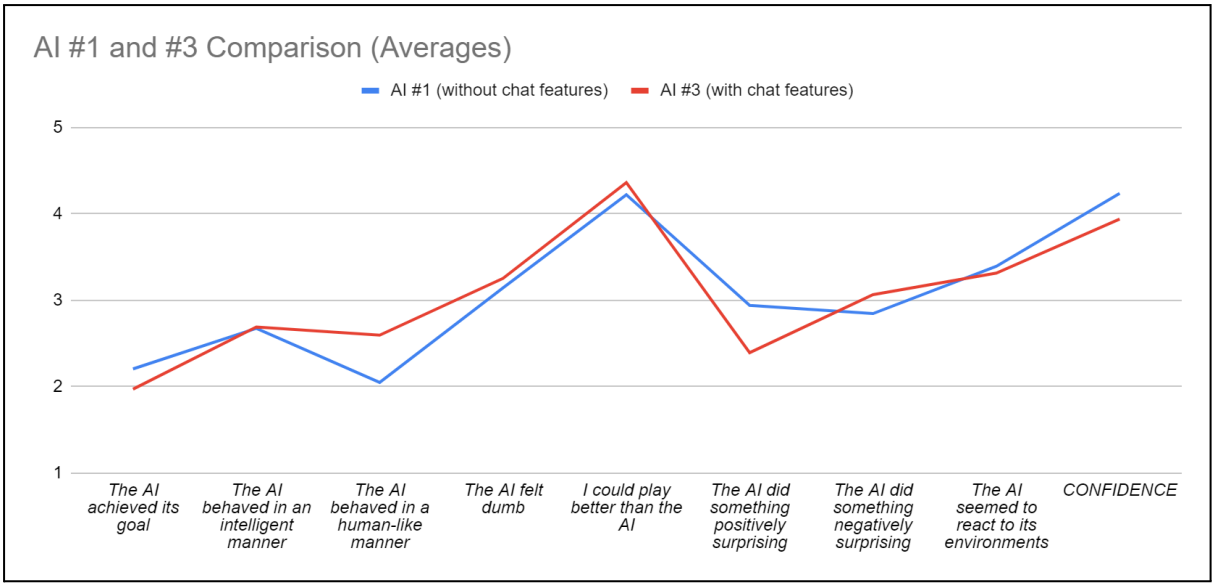
Appendix B: Survey Results

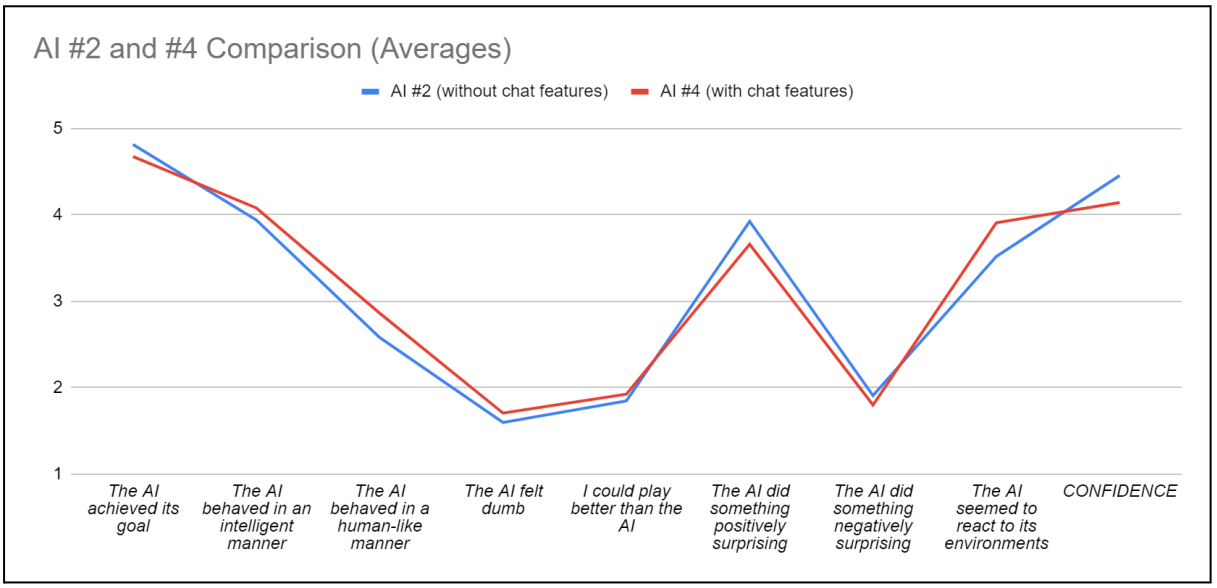
This appendix consists of the survey results collected for this thesis from January 20 to March 31, 2023, with some general figures about that data that were also used in the thesis. See the end of the appendix for the full data of the survey responses in CSV format.

Gender & Age Distribution

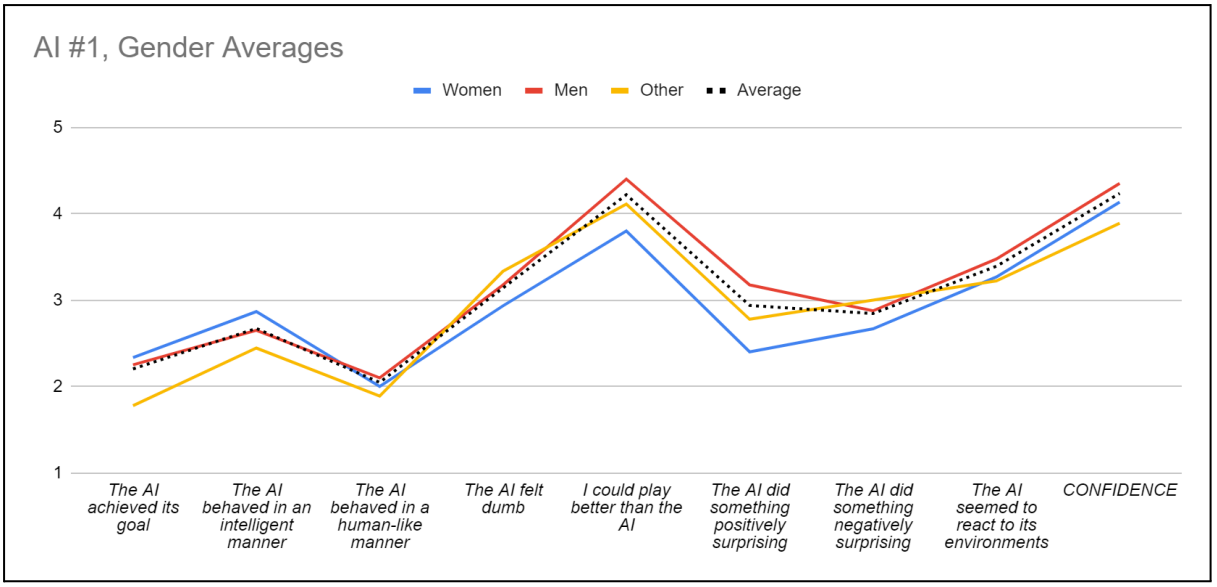


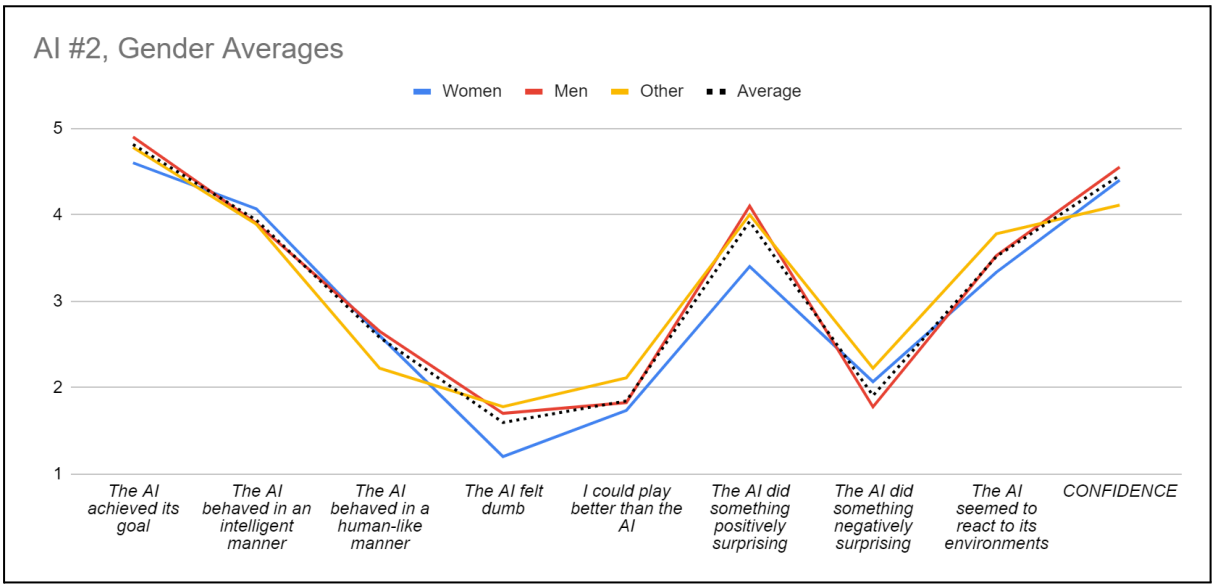
Overall Evaluation Averages

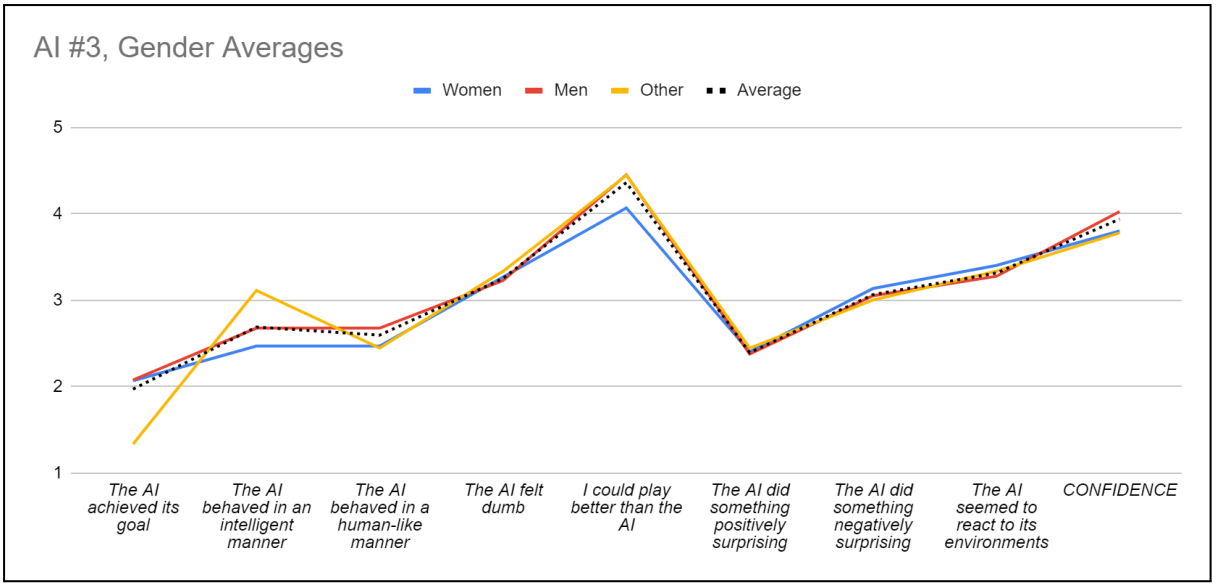


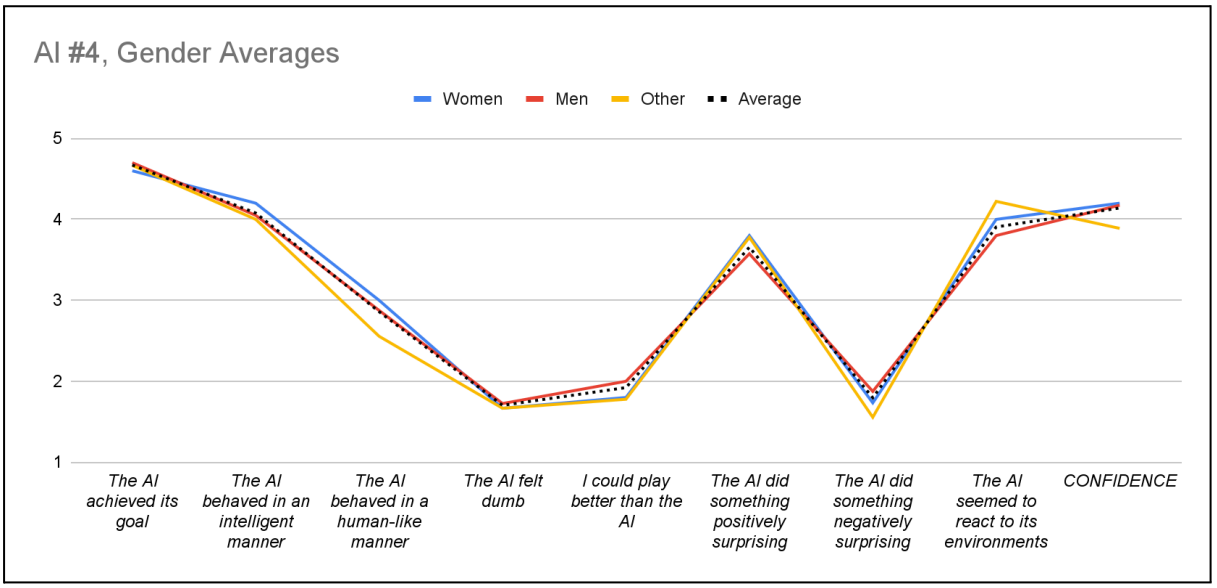


Evaluation Averages by Gender

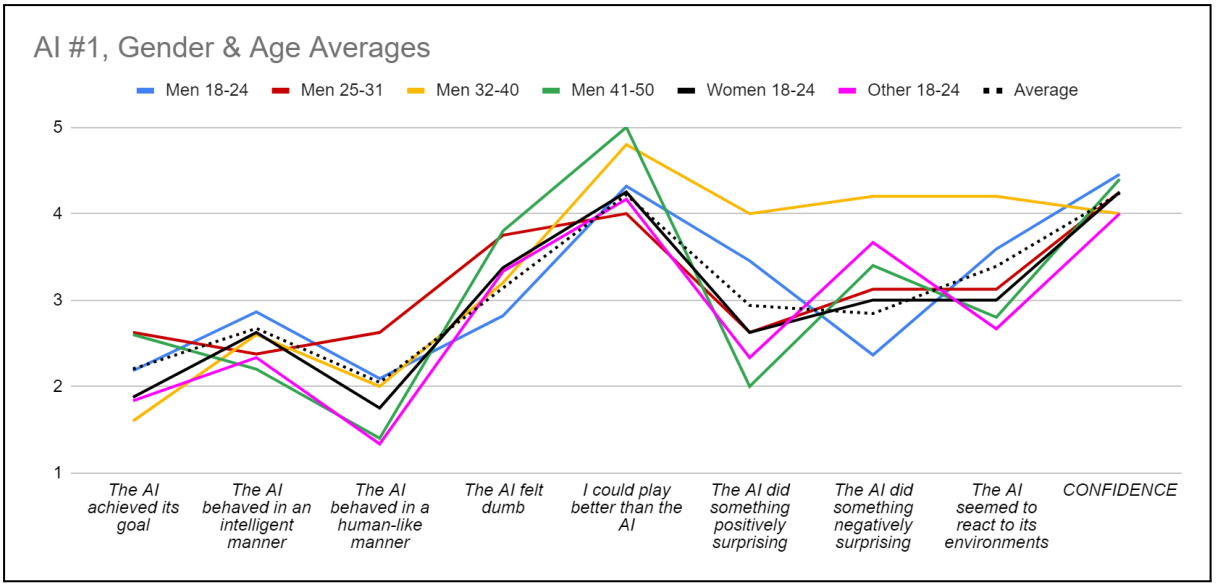


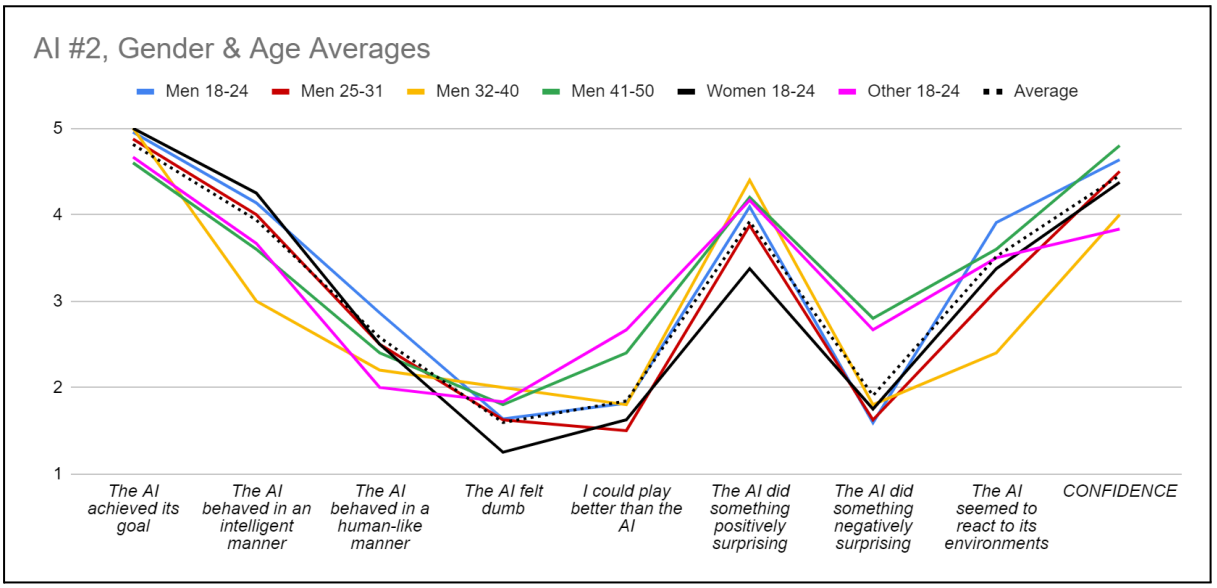


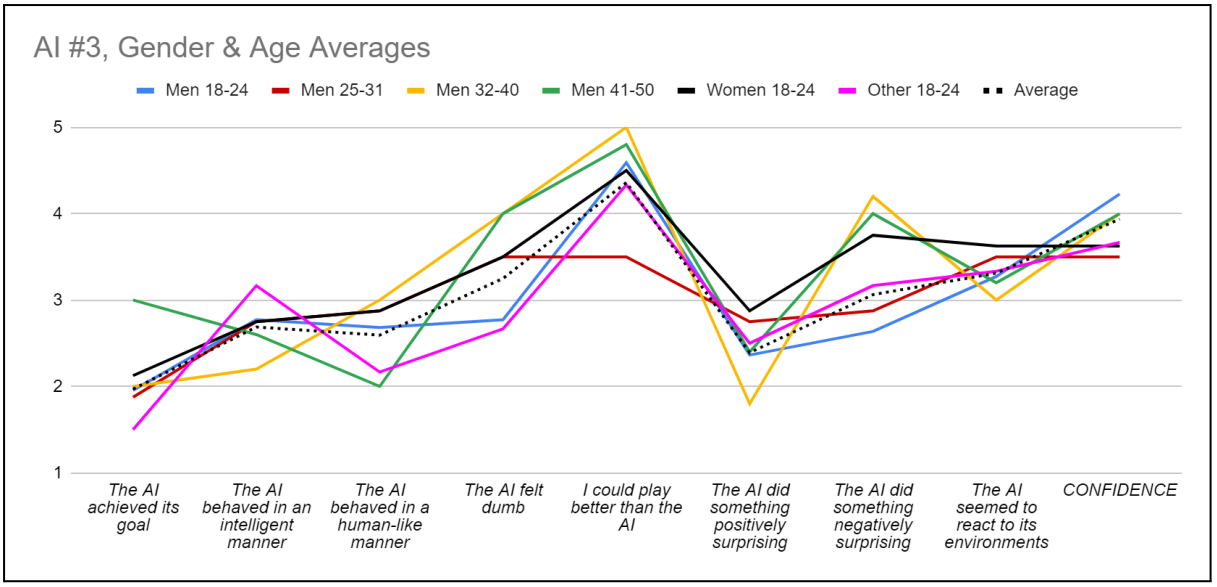

AI #3, Gender Averages
Women
Men
Other
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

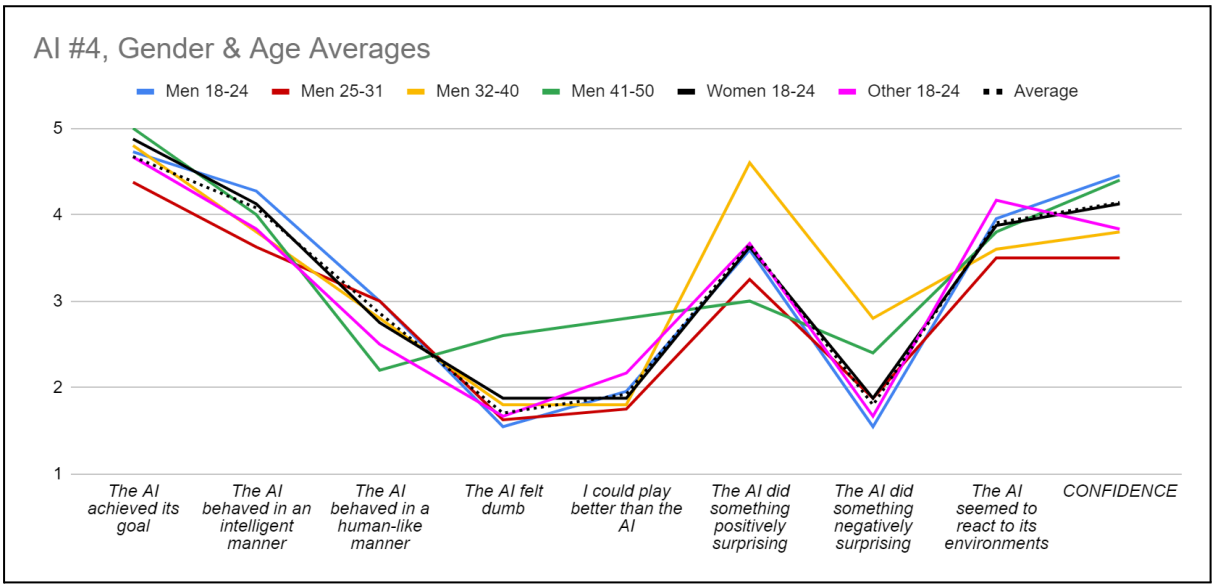

AI #4, Gender Averages
Women
Men
Other
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

Evaluation Averages by Gender & Age

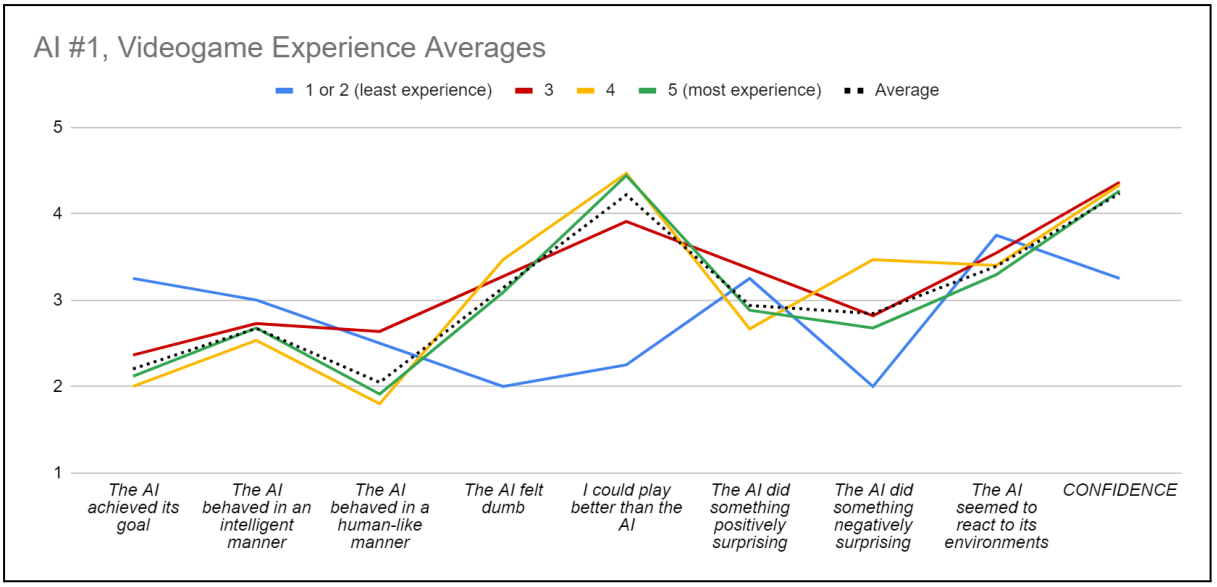


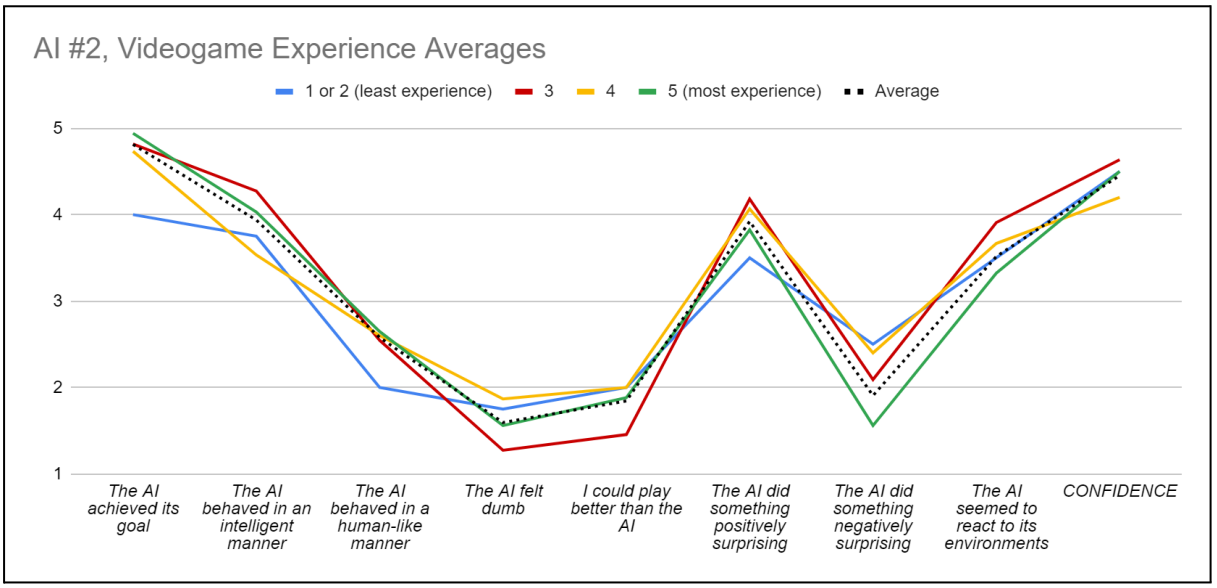


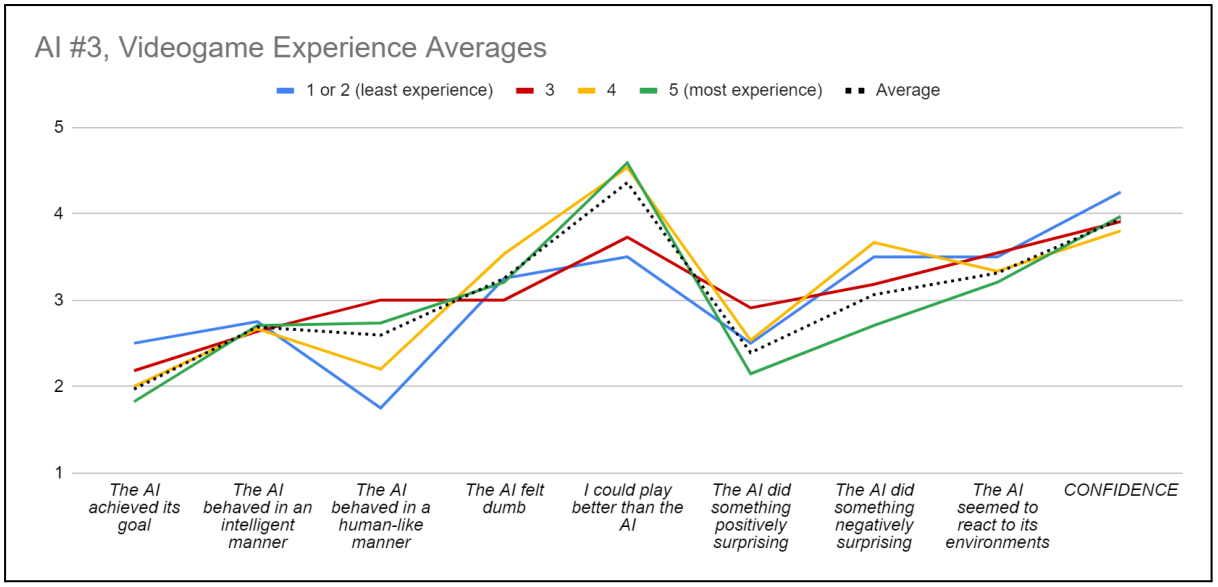

AI #3, Gender & Age Averages
Men 18-24
Men 25-31
Men 32-40
Men 41-50
Women 18-24
Other 18-24
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

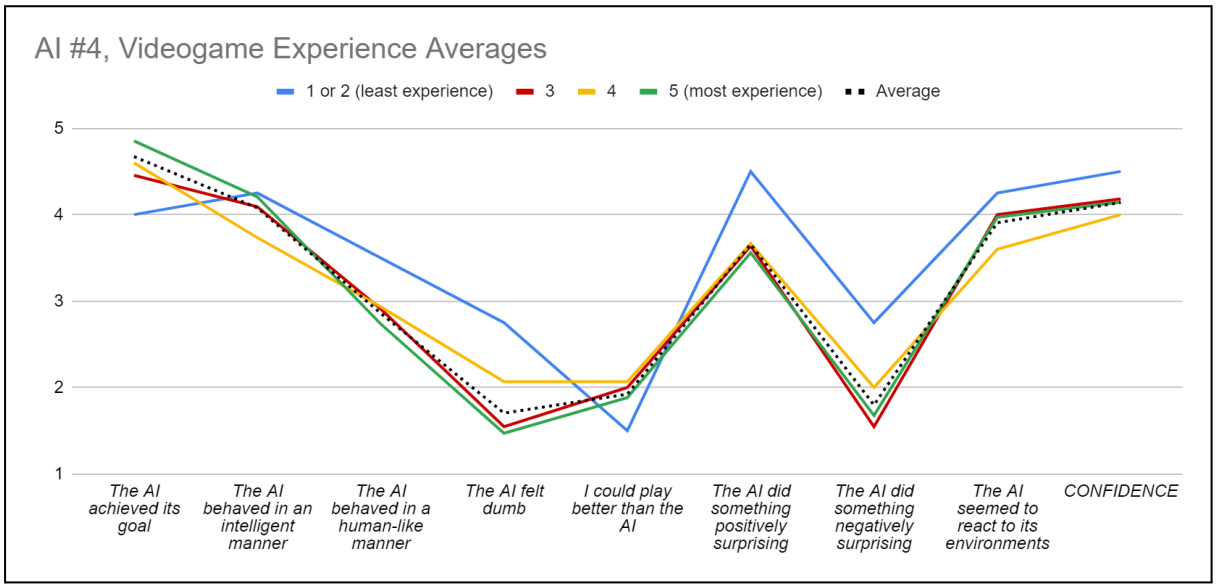

AI #4, Gender & Age Averages
Men 18-24
Men 25-31
Men 32-40
Men 41-50
Women 18-24
Other 18-24
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

Evaluation Averages by Videogame Experience

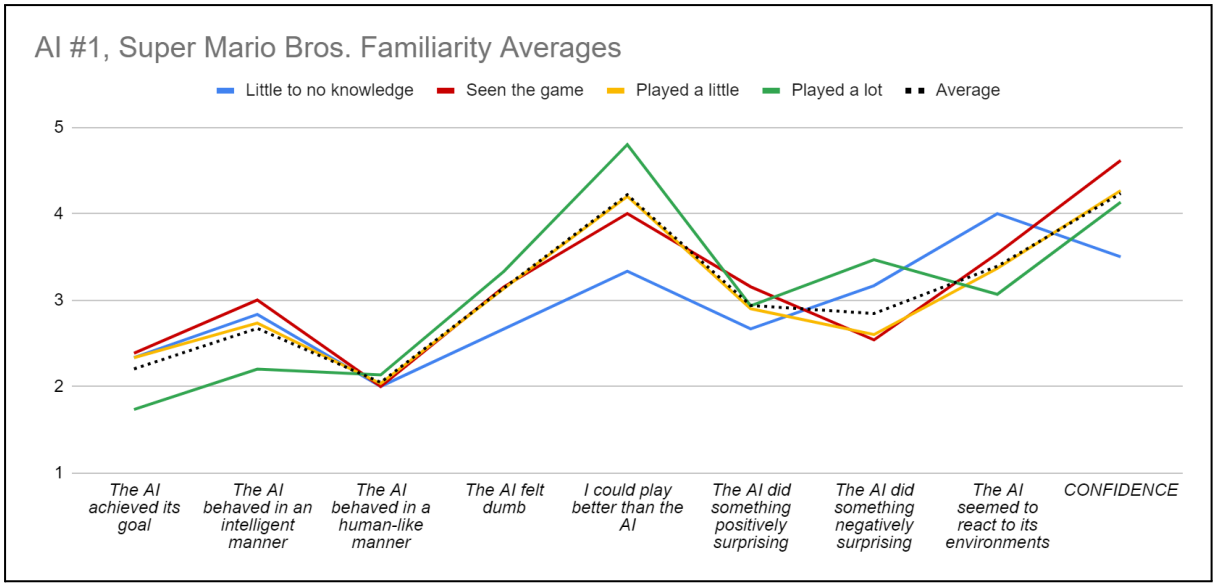


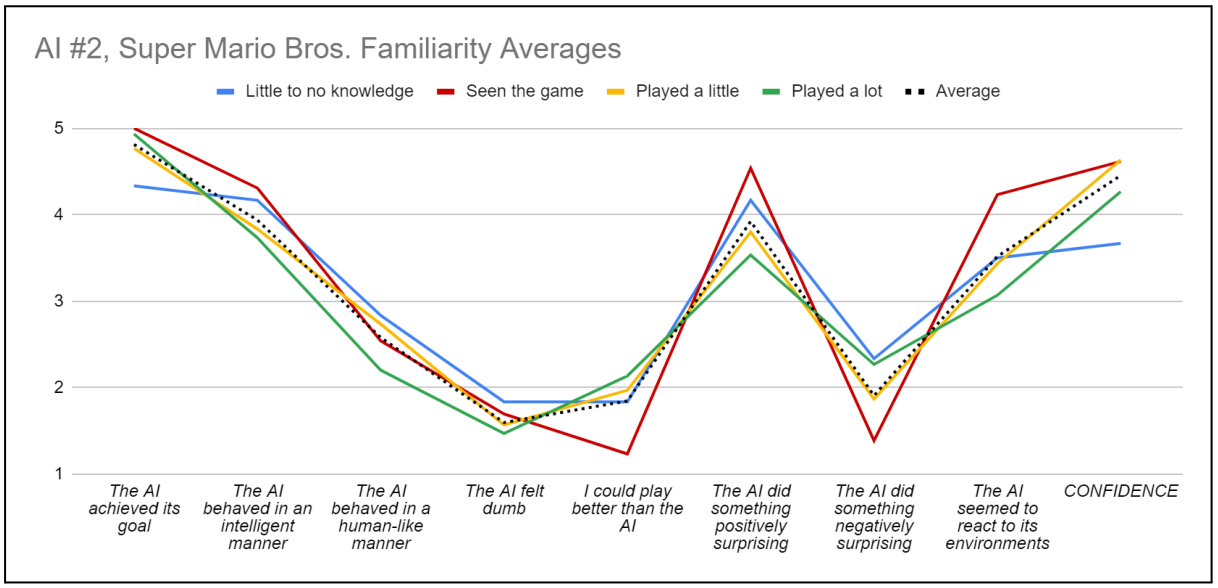


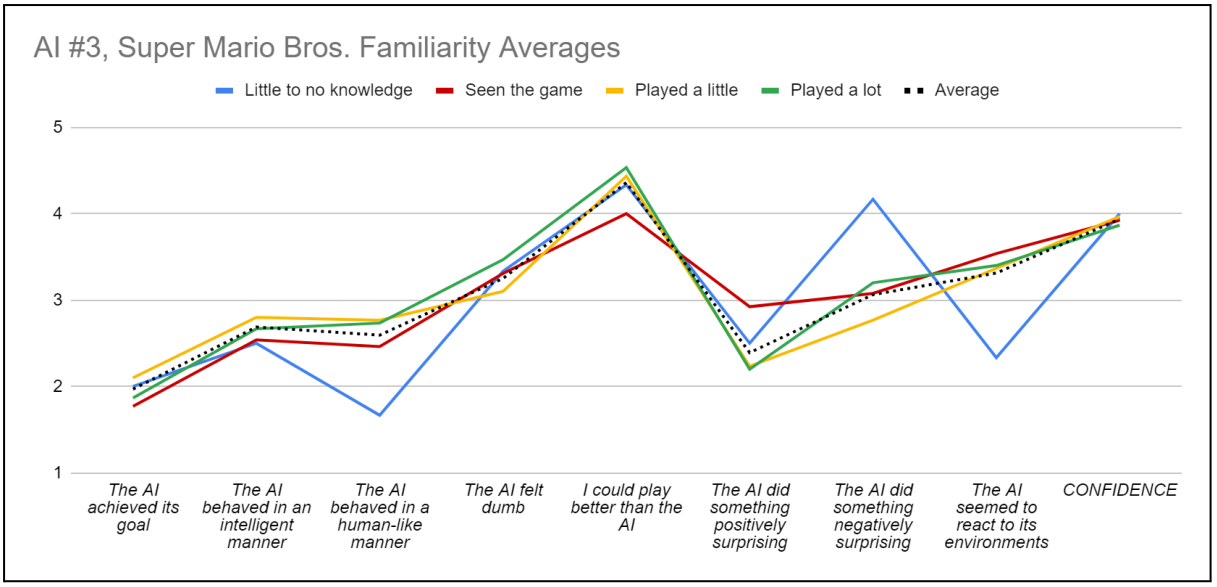

AI #3, Videogame Experience Averages
1 or 2 (least experience)
3
4
5 (most experience)
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

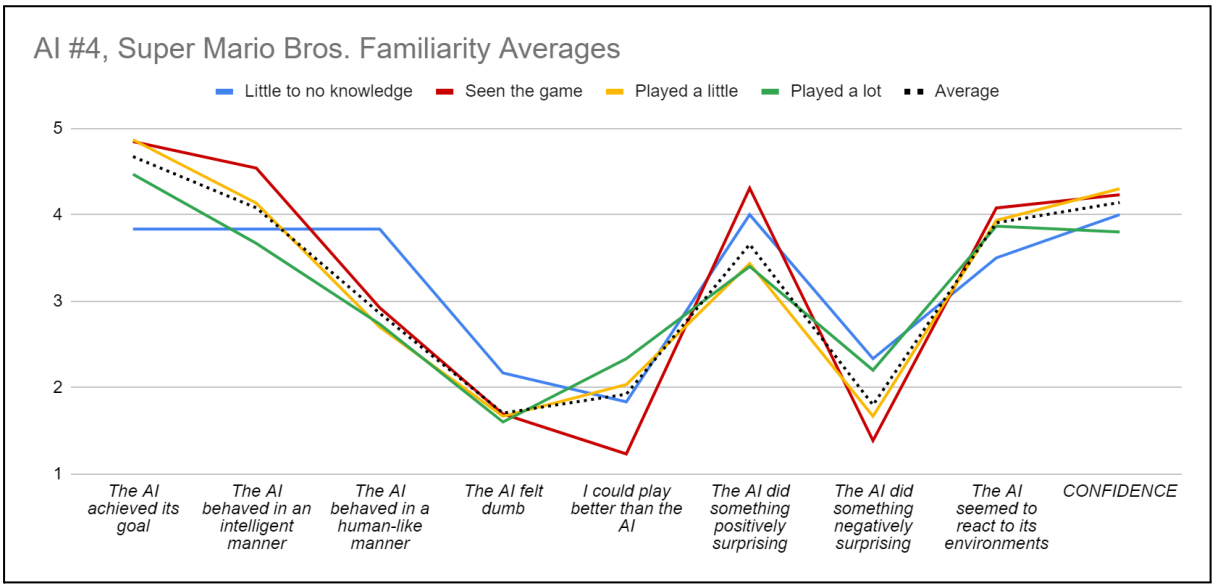

AI #4, Videogame Experience Averages
1 or 2 (least experience)
3
4
5 (most experience)
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

Evaluation Averages by *Super Mario Bros.* Familiarity

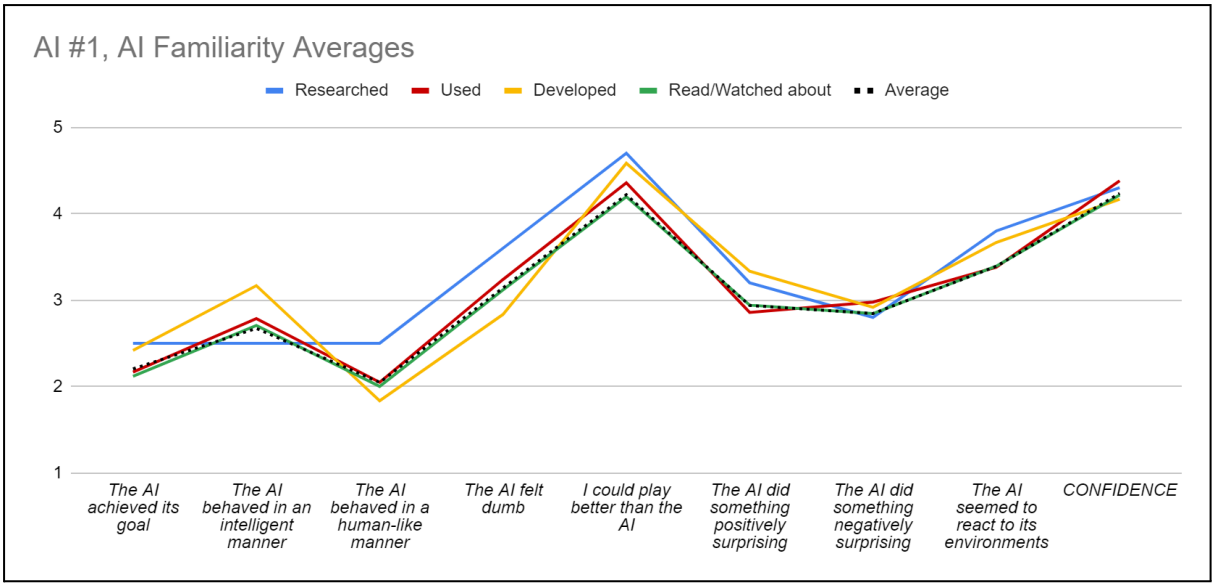


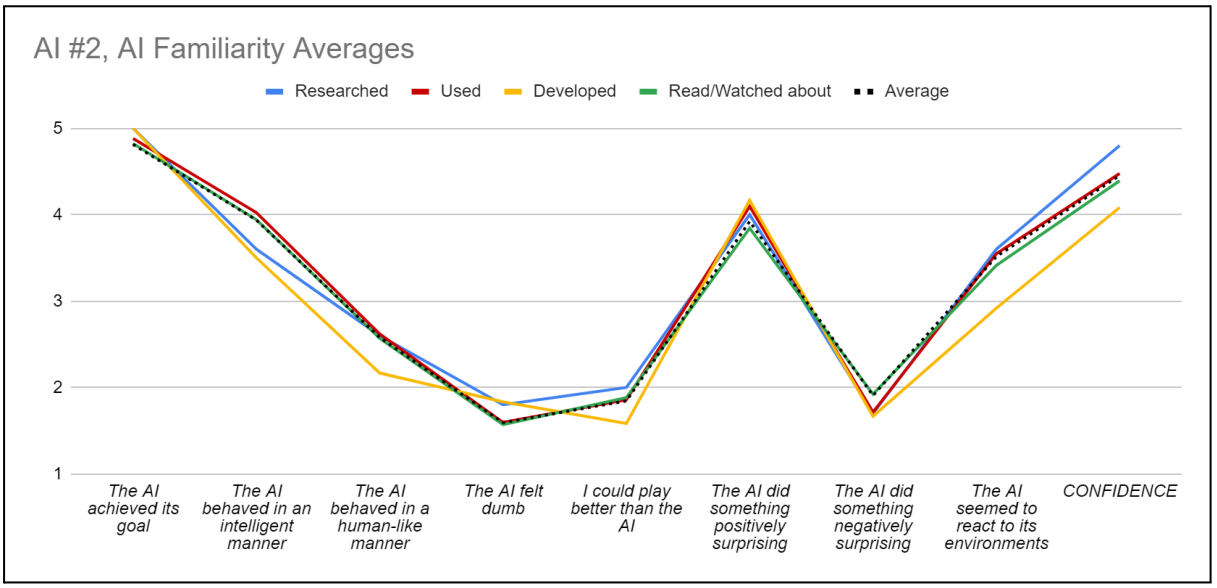


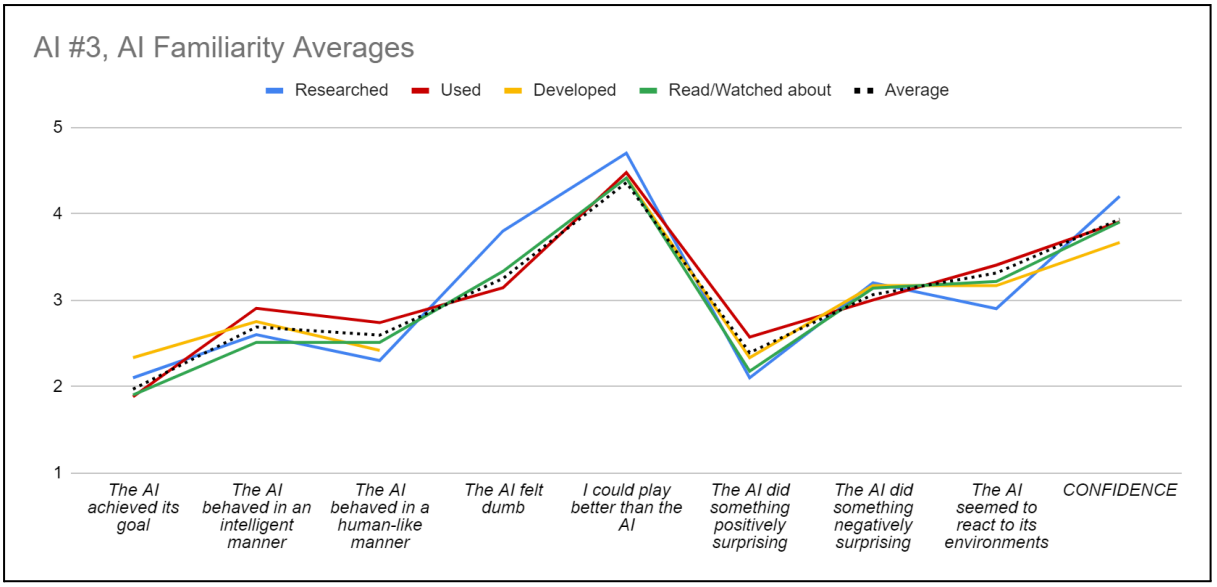

AI #3, Super Mario Bros. Familiarity Averages
Little to no knowledge
Seen the game
Played a little
Played a lot
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

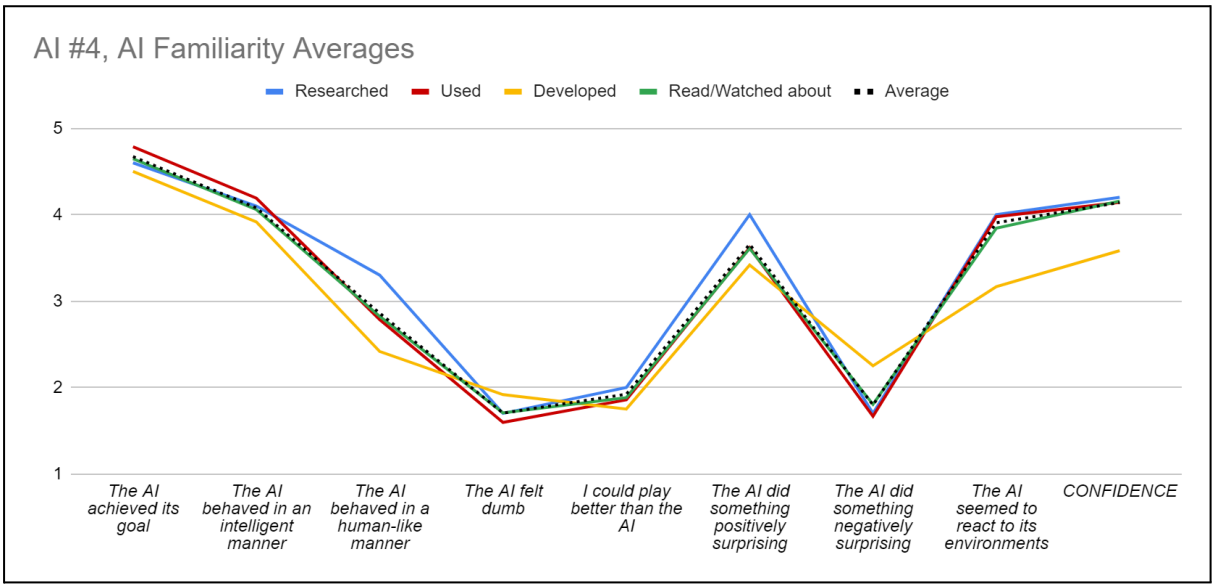

AI #4, Super Mario Bros. Familiarity Averages
Little to no knowledge
Seen the game
Played a little
Played a lot
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

Evaluation Averages by AI Familiarity

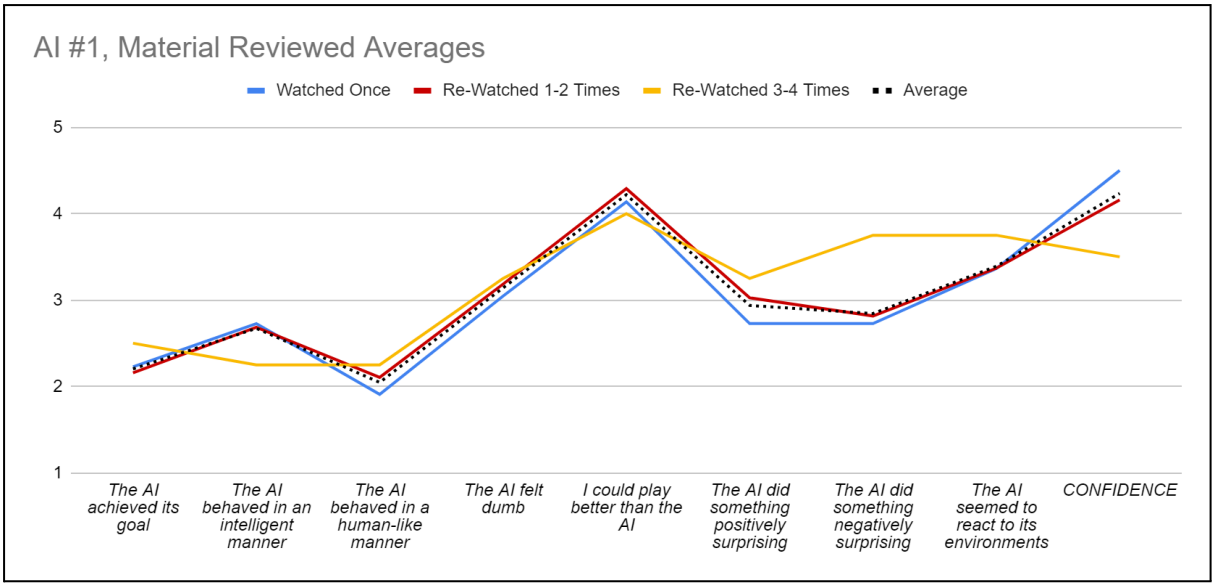


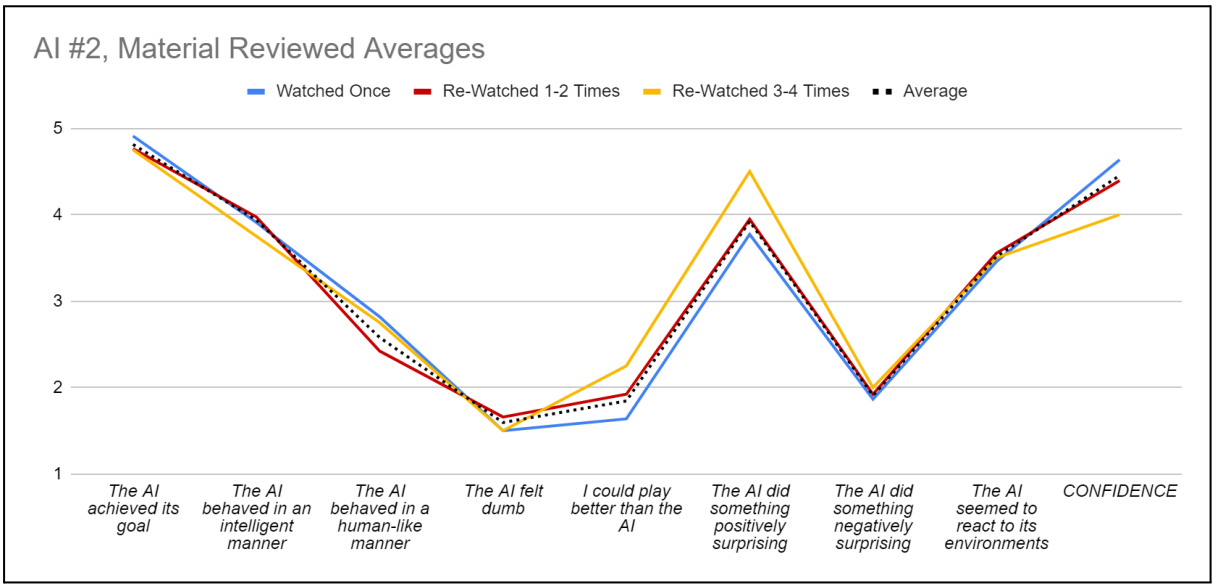


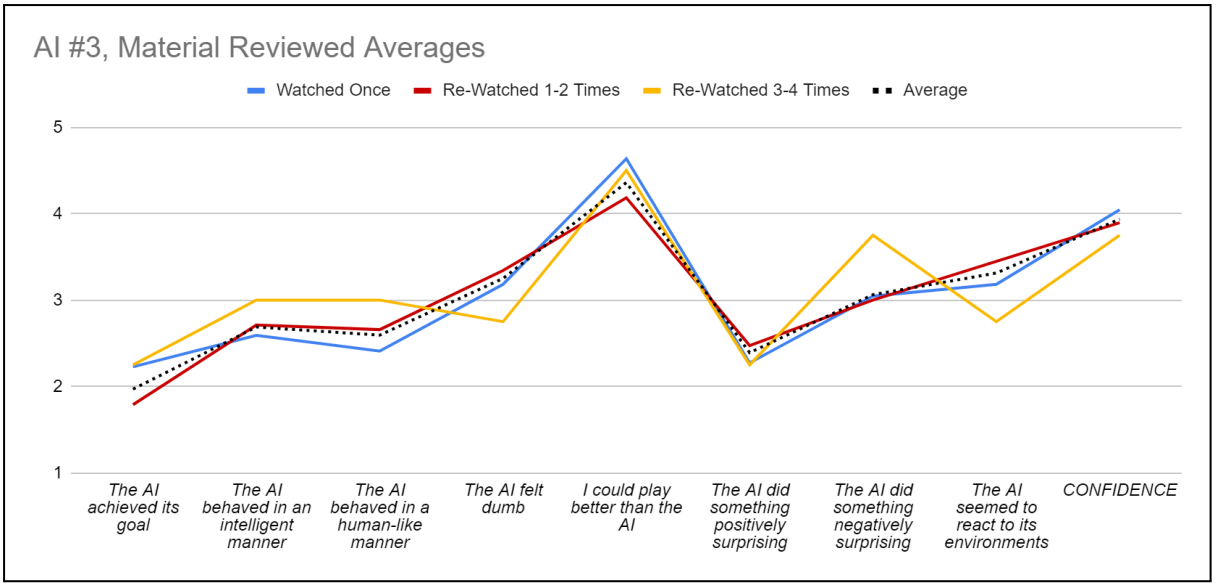

AI #3, AI Familiarity Averages
Researched
Used
Developed
Read/Watched about
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

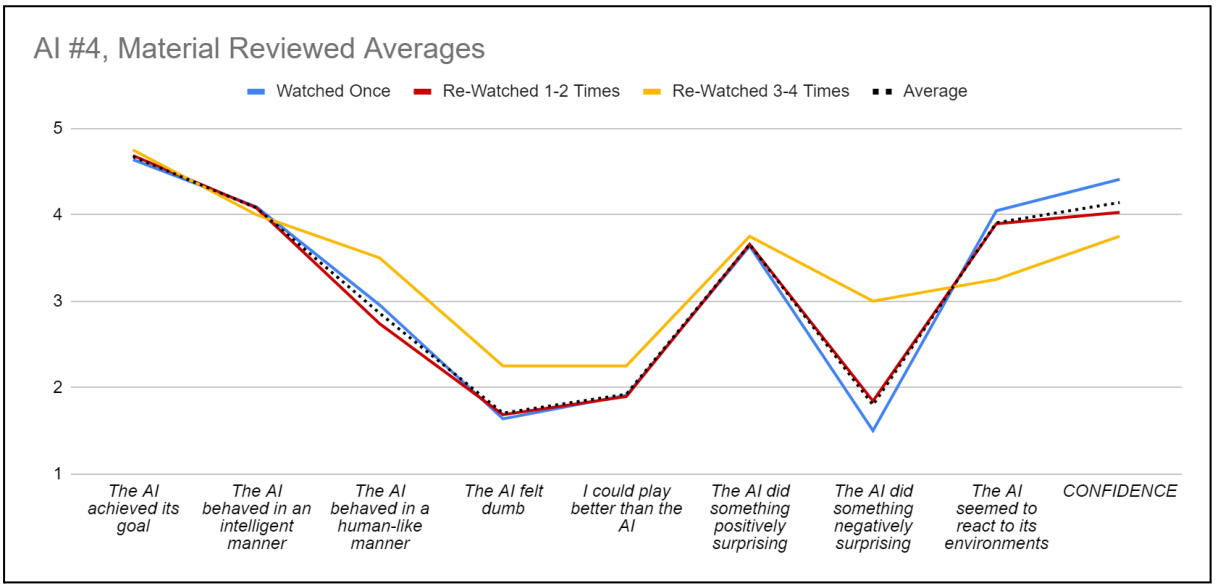

AI #4, AI Familiarity Averages
Researched
Used
Developed
Read/Watched about
Average
5
4
3
2
1
The AI achieved its goal
The AI behaved in an intelligent manner
The AI behaved in a human-like manner
The AI felt dumb
I could play better than the AI
The AI did something positively surprising
The AI did something negatively surprising
The AI seemed to react to its environments
CONFIDENCE

Evaluation Averages by Material Reviewed









All Survey Responses

The full data of the survey responses is provided below in CSV format, meaning that the tabular values are separated with a comma. It is possible to copy all the text in the next three pages to a text editor, save that as a .csv file, and then open the file in a spreadsheet software, such as Microsoft Excel, Google Sheets or Apache OpenOffice Calc. This data contains all the answers submitted to the survey, even the answer that was filtered out from the analysis.

Aikaleima,Age,Gender,"On a scale of 1 to 5, rate how experienced you are with videogames",How familiar are you with the original Super Mario Bros. videogame?,What is your familiarity with artificial
intelligence (AI)? You may select more than one option.,(Optional) Elaborate on your familiarity with Super Mario Bros. and with artificial intelligence.,The AI achieved its goal,The AI behaved in an
intelligent manner,The AI behaved in a human-like manner,The AI felt dumb,I could play better than the AI,The AI did something positively surprising,The AI did something negatively surprising,The AI
seemed to react to its environments,(Optional) How did evaluating the AI feel like?,"(Optional) Were there statements that you did not understand, or had a hard time answering to?","On average, how
confident were you in your answers?",The AI achieved its goal,The AI behaved in an intelligent manner,The AI behaved in a human-like manner,The AI felt dumb,I could play better than the AI,The AI did
something positively surprising,The AI did something negatively surprising,The AI seemed to react to its environments,(Optional) How did evaluating the AI feel like?,"(Optional) Were there statements
that you did not understand, or had a hard time answering to?","On average, how confident were you in your answers?",The AI achieved its goal,The AI behaved in an intelligent manner,The AI behaved
in a human-like manner,The AI felt dumb,I could play better than the AI,The AI did something positively surprising,The AI did something negatively surprising,The AI seemed to react to its
environments,(Optional) How did evaluating the AI feel like?,"(Optional) Were there statements that you did not understand, or had a hard time answering to?","On average, how confident were you in
your answers?",The AI achieved its goal,The AI behaved in an intelligent manner,The AI behaved in a human-like manner,The AI felt dumb,I could play better than the AI,The AI did something positively
surprising,The AI did something negatively surprising,The AI seemed to react to its environments,(Optional) How did evaluating the AI feel like?,"(Optional) Were there statements that you did not
understand, or had a hard time answering to?","On average, how confident were you in your answers?","For each evaluation, on average, how many times did you return to the video and/or the
screenshot, when answering the questions?","(Optional) Feel free to leave any additional information or last thoughts about AI, Mario, and the survey here!"
23.1.2023 klo 8.31.14,41 - 50,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,1,3,1,4,5,1,4,3,Simple,No,4,4,4,3,2,3,4,2,4,Simple,No,5,3,3,3,3,4,2,3,3,Simple,a bit,3,5,4,3,1,2,4,2,5,Simple,No,5,Re-checked the material 1 - 2 times,"Interesting. When GPT4 begins to create new AI by
itself, we are in interesting situation. Is there something AI 'likes' to do? for example gaming or other 'fun' without any other purpose? Is it possible that AI begins to play with humans?"
23.1.2023 klo 8.39.06,25 - 31,Male,4,I have played the game a little bit,I have read or watched about AI,,2,3,2,4,5,2,3,2,,,5,5,4,4,1,1,4,3,3,,,5,3,3,3,3,3,3,3,3,,,3,4,3,3,3,1,3,3,3,,,4,"Only watched the material
once, did not return to it",
23.1.2023 klo 9.30.01,18 - 24,Female,3,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,2,4,1,4,4,4,2,4,,,4,5,4,3,1,1,5,1,3,,,4,2,4,4,4,4,4,3,4,,,3,5,5,3,1,1,5,1,4,,,4,Re-checked the material 1 - 2 times,
23.1.2023 klo 9.31.19,32 - 40,Male,5,I have played the game extensively,I have read or watched about AI,,1,2,1,2,5,5,5,4,,,4,5,4,1,1,1,5,2,2,,,4,1,2,2,4,5,2,5,4,,,4,5,4,1,1,1,5,4,4,,,4,Re-checked the material 1 - 2
times,"In the first video with chat interface, the AI's answers seem like ""I did it because I was programmed to do so"", which doesn't feel like intelligence but a repeatable predetermined routine."
23.1.2023 klo 9.42.32,18 - 24,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI
techniques, I have read or watched about AI",,4,4,1,2,5,4,1,3,,,4,5,2,4,2,2,5,2,1,,"It's hard to speak for it's intelligence because it is almost as if it is jumping pixel perfect jumps, as if someone gave it a map
beforehand that said the exact spots to jump for the fastest performance but this also may be the conclusion that AI arrives to when given the goal of completing the level as fast as
possible",4,5,2,2,2,5,2,4,5,,,4,5,5,3,2,2,5,2,5,,,5,"Only watched the material once, did not return to it","It has been a while since I played Mario but in my mind the most human-like behavior one could do is
jump for the question mark tiles, they give rewards, don't they? Also, human curiosity of seeing a question mark would prompt almost any human to interact with it."
23.1.2023 klo 9.50.53,18 - 24,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,2,2,2,4,5,3,2,4,It was difficult to
evaluate the AI because not much information was given about the AI. Such as did the AI react to the hidden block it hit or was it just the random backtrack it performed through the run.
,,4,5,4,4,1,3,5,1,4,"Evaluating the AI felt significantly easier than on the previous evaluation. Mainly because now the run felt more like a TAS (tool-assisted) run than a ""dumb"" AI
run.",,5,1,2,4,2,5,1,1,2,,,4,5,4,3,1,4,1,1,4,,,4,Re-checked the material 3 - 4 times,
23.1.2023 klo 9.57.52,18 - 24,Male,3,I have played the game extensively,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched
about AI",,2,3,4,3,5,4,3,4,,,4,5,5,2,1,2,5,1,5,,,5,2,2,3,4,5,4,4,4,,,4,5,5,2,1,2,5,1,5,,,5,Re-checked the material 1 - 2 times,
23.1.2023 klo 9.59.33,18 - 24,Male,5,I have seen videos of someone else playing the game,I have used a program that utilizes AI techniques (e.g.
ChatGPT),,2,3,3,3,5,3,3,3,,,5,5,5,3,1,3,5,1,5,,,5,3,3,3,3,3,3,3,3,,,5,5,5,2,1,2,5,2,5,,,5,Re-checked the material 1 - 2 times,
23.1.2023 klo 10.02.29,41 - 50,Male,3,I have played the game a little bit,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",I was a teen in the 90's. I know Nintendo games. In UTU I've participated to multi-diciplinary course (25 cr) about AI and I'm utilizing it at work.
,3,2,3,4,5,2,2,3,Neutral,No,5,5,2,3,3,4,4,2,3,Neutral,No,5,2,3,1,4,5,2,5,3,,,5,5,4,3,3,4,1,2,4,,,5,Re-checked the material 1 - 2 times,
23.1.2023 klo 10.18.32,32 - 40,Male,5,I have played the game extensively,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped
with creating a program that utilizes AI techniques, I have read or watched about AI",I grew up playing Super Mario Bros and one of my main research interests is AI. ,3,3,4,3,5,5,5,5,"It felt like watching
someone play the game, but there were subtle things in the movement that felt off, a tad like the uncanny valley. It appears to have the goal of reaching the end of the level as quickly as possible.
However, it prioritizes this over certain behaviors which enable recovery from mistakes (not getting the mushroom). I made mistakes that were based of the programming as well. Particularly the
moment at 00:19 in which it backs up only to fall into the space between hills. It appears that it was priming for a jump, but given that the downslope does not end in a pit it was unnecessary. ","#1, I
do not know the goal of the AI. It did not acheive the goal of the game (to reach the end of the level) or the goal of a person (to play for enjoyment). ",4,5,3,2,2,2,4,2,3,"It felt like watching a speed
runner. There was no pausing, no stopping, just an immediate sprint towards the end. However, it takes time to trigger a coin and a mushroom, the latter which it does not get. ",No.
,5,3,4,4,3,5,2,4,3,"Like watching a new person play. They don't go after the question boxes, they don't stomp goombas for points.","Yes. When a question was asked why it jumped earlier. It used the
same timestamp, but gave a different pattern type. It couldn't be the same jump. It could be the action sequence directly before the jump at the timestamp, but then why the same time stamp.
",4,5,4,4,2,1,5,4,4,"Again, like a speed runner. The chatbot gives the appearance of uncertainty while the chatbot gives direct answers. This estimation is interesting. I am interested in why there are so
many messages with the same timestamp. I get that they trigger from the same instance. I would recommend putting the timer on the screen. ",No. ,4,Re-checked the material 3 - 4 times,"Great job!
The affective Mario chat based on the generalization of situations is pretty good. Notably, it seems more human when achieving a speed run task. The main obstacle to it feeling human-like is that it
does not explore or seek out items. It seems that sub-goals are necessary to achieve this. With human play, priorities shift. For instance, the survival priority can be lowered in situations with items that
a player wants and thus will risk a life to get. This, of course, is contingent on the number of remaining lives. There is also a matter of points (which I never went after) and coins, which may reduce
situational survivability but increase long-term survivability. I would deem this an effective and interesting project. Best of luck!"
23.1.2023 klo 10.19.15,18 - 24,Male,5,I have played the game a little bit,"I have done research involving AI, I have read or watched about
AI",,2,2,1,4,5,4,4,5,,,5,5,5,5,1,3,5,1,5,,,5,3,2,2,4,5,3,4,3,,,5,5,5,5,1,3,5,1,5,,,5,"Only watched the material once, did not return to it",
23.1.2023 klo 10.38.23,18 - 24,Male,3,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,2,2,2,4,4,4,3,2,,,4,5,5,4,1,1,5,1,5,,,5,2,2,3,2,4,2,2,4,,,4,5,5,4,1,1,5,1,5,,,4,Re-checked the material 1 - 2 times,
23.1.2023 klo 11.31.06,18 - 24,Male,4,I have seen videos of someone else playing the game,I have read or watched about AI,,5,3,1,4,5,4,4,4,,,4,1,5,1,1,5,5,1,5,,,5,5,5,5,1,1,5,5,5,,,2,2,5,3,5,5,5,5,5,,,5,"Only
watched the material once, did not return to it",
23.1.2023 klo 11.45.12,25 - 31,Male,2,I have seen videos of someone else playing the game,I have read or watched about AI,,4,3,3,3,1,4,2,4,Good,No,4,5,3,1,3,1,5,1,5,No
feelings,No,5,2,3,3,3,1,3,3,4,Maybe Im only a algorithm,NO,4,5,3,3,3,1,3,3,4,"Thinking how similar humans and AI are? Do we overestimate humans or underestimate AI? Humans have creativity and
intuition, but humans use also rules of thumb, simple locig etc.",No,4,Re-checked the material 1 - 2 times,Mario is italian
23.1.2023 klo 12.11.08,18 - 24,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,4,3,2,3,3,4,4,3,,,5,5,5,4,1,3,4,1,5,,,5,4,3,4,2,4,2,1,4,,,5,5,4,2,1,3,2,1,3,,,5,"Only watched the material once, did not return to it",
23.1.2023 klo 12.37.08,18 - 24,Male,4,I have seen videos of someone else playing the game,"I have done research involving AI, I have read or watched about
AI",,2,3,2,3,5,4,2,4,,,4,5,5,2,1,1,5,2,5,,,5,2,2,2,4,5,2,3,3,,,3,5,5,4,1,1,4,1,5,,,4,"Only watched the material once, did not return to it",
23.1.2023 klo 14.09.19,18 - 24,Male,4,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes
AI techniques, I have read or watched about AI, I study AI at university",,1,5,3,1,5,3,3,5,,"""The AI achieved its goal"" depends on the goal. It didn't finish the level, but it did play quite
well.",4,5,5,3,1,1,3,3,5,,"""The AI achieved its goal"" Depends on the goal ",4,1,5,3,1,5,3,3,5,,,4,5,5,3,1,1,3,3,4,,,4,Re-checked the material 1 - 2 times,
23.1.2023 klo 14.10.00,18 - 24,Female,4,I have played the game extensively,I have read or watched about AI,,1,2,2,2,5,2,4,3,,,4,5,2,2,1,4,2,3,3,,,4,1,1,1,5,5,2,4,4,,,4,5,2,3,3,4,4,3,4,,,4,Re-checked the material 1
- 2 times,
23.1.2023 klo 14.46.20,41 - 50,Male,4,I have played the game extensively,I have read or watched about AI,,2,1,1,5,5,1,5,3,,,4,5,4,4,1,3,3,3,4,,,4,5,1,1,5,5,1,5,3,,,4,5,4,3,2,3,3,3,3,,,4,"Only watched the material
once, did not return to it",
23.1.2023 klo 16.12.36,18 - 24,Male,5,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI","Im a fan
of platformer games in general, have played other mario titles and seen videos and speedruns of original game aswell. ",2,3,1,2,5,5,2,1,neutral,no,5,5,5,2,1,1,5,1,4,,again
no.,5,1,3,1,2,4,4,2,5,,,5,5,5,2,1,1,5,1,5,,,4,"Only watched the material once, did not return to it",
23.1.2023 klo 16.20.05,18 - 24,Male,4,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a
program that utilizes AI techniques, I have read or watched about AI",,2,3,1,3,2,3,2,3,"Evaluating the AI felt easy, there is clearly room for improvement. For example the backtracking of the AI and getting
stuck for few seconds is something that experienced player wouldn't do.",,5,5,2,2,4,1,4,1,3,Evaluating the AI became harder since there isn't any clear mistakes the model does. Evaluating the
intelligence of the AI is hard since you can't really be sure whether or not the AI reacts to the environment or that is has just learned obstacles of the level.,,4,2,3,2,4,4,3,3,2,,,4,4,3,2,3,1,4,2,2,,,4,"Only
watched the material once, did not return to it",Seems like the overall view of the AI is more intelligent when the AI comes up with reasons for it's behaviour.
23.1.2023 klo 16.48.17,18 - 24,Male,5,I have seen videos of someone else playing the game,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have
read or watched about AI",,4,4,3,2,2,4,2,5,Just fine,No they were very simple and clear,5,5,3,1,2,1,3,3,3,just fine again,no,5,2,3,1,4,5,2,4,3,good,no,4,5,5,3,1,1,5,1,4,,,5,"Only watched the material once, did
not return to it",
23.1.2023 klo 18.42.35,18 - 24,Male,5,"I have seen/played other Mario games, but I have not seen or played Super Mario Bros.","I have used a program that utilizes AI techniques (e.g. ChatGPT), I have
read or watched about AI",,2,3,2,2,5,4,2,4,,,4,5,5,2,2,3,3,3,2,,,4,1,3,2,2,5,3,3,1,,,4,5,3,3,2,3,3,3,2,,,4,Re-checked the material 1 - 2 times,The AI does not seem to react to the environment that much. It is more
or less set on a path and will do the required inputs to achieve the end goal.
23.1.2023 klo 20.07.05,25 - 31,Male,5,I have played the game extensively,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped
with creating a program that utilizes AI techniques, I have read or watched about AI",,4,3,1,4,5,3,3,2,"Seems like this version reward is too biased to siomply move forward, i think adding a small
negative reward every time step would help it optimize the traversal towards the end",,4,5,5,1,1,2,4,2,2,"This seems like a genetic algorithm that has just found the right sequence of inputs, and probably
would not generalize well to other level or random levels",,4,2,2,2,5,2,2,2,2,,,2,2,2,2,2,2,2,2,2,I dont really understand the chat feature,,1,Re-checked the material 1 - 2 times,Cool! I'm also working on a RL
agent for RTS
23.1.2023 klo 20.32.12,18 - 24,Male,5,I have played the game a little bit,I have read or watched about AI,,3,2,1,3,5,1,1,1,,,4,5,4,2,4,3,1,1,1,,,4,2,1,3,3,5,1,1,2,,,4,5,3,3,3,3,1,1,1,,,4,Re-checked the material 1 - 2
times,"The ""good"" AI still feels dumb because while it runs through the level near perfectly, it does not incorporate basic tricks like taking the pipe or landing near the base of the flag for a faster end
animation."
23.1.2023 klo 23.50.02,18 - 24,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,2,3,1,2,4,4,1,2,,,5,5,3,2,1,2,4,1,3,,,5,2,4,4,3,5,2,2,5,,,4,5,3,1,2,1,3,1,2,,,5,Re-checked the material 1 - 2 times,
24.1.2023 klo 0.26.22,18 - 24,Male,5,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",,2,1,1,4,5,2,3,4,,,4,5,4,2,3,1,4,1,5,,,4,1,2,2,4,5,4,3,4,,,3,5,4,2,4,1,4,1,4,,,4,Re-checked the material 1 - 2 times,
24.1.2023 klo 1.01.34,18 - 24,Female,3,I have seen videos of someone else playing the game,I have read or watched about
AI,,3,4,4,3,4,5,1,4,,,5,5,5,2,2,1,4,2,3,,,4,3,2,4,3,3,3,3,4,,,3,4,4,2,2,1,4,2,3,,,3,Re-checked the material 1 - 2 times,
24.1.2023 klo 3.19.50,25 - 31,Male,4,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI
techniques",,2,2,2,4,5,3,3,3,,,4,5,3,1,2,1,4,1,1,,"The AI behaved in an intelligent manner
a complex question on a complex issue",3,3,3,1,4,4,3,3,4,,,2,4,3,1,1,1,1,1,1,,,2,Re-checked the material 1 - 2 times,Cool project
24.1.2023 klo 12.45.22,18 - 24,Male,5,I have played the game a little bit,I have created or helped with creating a program that utilizes AI
techniques,,2,2,2,2,4,4,2,4,,,4,5,2,3,2,1,4,1,4,,,5,2,3,3,2,4,2,2,3,,,4,5,5,4,1,1,4,1,4,,,4,Re-checked the material 1 - 2 times,
24.1.2023 klo 23.43.08,18 - 24,Male,5,I have played the game a little bit,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about
AI",I have played most mario games on Wii. Very familiar with the franchise and platformers in general from the content I view and speedrunning. AI is a big in the tech industry so I like to read about it
when ever possible / some1 makes some cool shit.,1,2,3,5,5,1,3,4,"Similar to evaluating a human performing. The task here is of course decently limited, which makes evaluating easier.","We haven't

been told what the goal of the AI was... Makes the 1st question confusing. We are forced to assume the AI's goal was to finish the stage, since the title states ""play THROUGH"",5,5,3,5,1,1,5,1,5,Same as before.,Again forced to assume that the AI's goal is to finish the level.,5,1,5,5,1,5,1,1,5,Same as before,Still assuming the goal is to finish the level.,5,5,5,5,1,1,5,1,5,Same as before,Same as before,5,"Only watched the material once, did not return to it",
25.1.2023 klo 1.59.49,18 - 24,Female,2,"I have not seen/played other Mario games, but I have not seen or played Super Mario Bros.",I have read or watched about AI,,2,2,2,2,2,4,4,4,,,4,5,5,2,1,1,5,1,4,,,4,4,4,2,2,4,4,5,4,,,4,5,4,4,4,1,5,4,3,,,4,Re-checked the material 3 - 4 times,
25.1.2023 klo 23.05.11,18 - 24,Female,3,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,1,2,2,4,5,1,5,3,Easy,,5,5,5,5,5,1,2,5,2,5,Easy,,5,1,2,3,3,5,4,5,4,Easy,,5,5,5,4,1,3,5,1,5,Easy,,5,Re-checked the material 1 - 2 times,The AI seemed to be doing a speedrun and not to collect coins and boosters on the way.
25.1.2023 klo 23.46.00,18 - 24,Female,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,2,2,1,4,4,2,1,2,,,3,5,5,1,2,1,1,1,4,,,5,2,2,4,4,5,2,4,2,,,2,5,4,1,1,1,1,1,4,,,5,"Only watched the material once, did not return to it",
26.1.2023 klo 14.07.26,18 - 24,Male,4,"I have seen/played other Mario games, but I have not seen or played Super Mario Bros.","I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,1,3,3,4,4,4,4,5,,,5,4,5,3,2,2,5,2,4,,,4,1,1,1,5,5,1,4,2,,,5,4,4,4,2,2,3,2,4,,,4,Re-checked the material 1 - 2 times,
27.1.2023 klo 14.53.39,32 - 40,Male,4,I have played the game extensively,I have read or watched about AI,,1,2,2,4,5,4,4,4,,"The intelligent of AI. It seemed to dodge the obstacles and enemies but didn't collect mushrooms, coins etc. So is it intelligent? Kind of when trying not get killed but not when you think about collecting coins and stuff that help you advance in the game. Also the AI seemed to react to it's environment regarding obstacles and enemies but not when considering collecting stuff. ",4,5,2,3,3,2,4,2,4,,,4,2,2,4,4,5,2,4,4,,,4,5,4,4,2,2,4,2,4,,,4,Re-checked the material 1 - 2 times,Best of luck with your research!
28.1.2023 klo 9.08.04,41 - 50,Female,1,I barely know anything about Mario,I have read or watched about AI,"I played a Super Mario game a few times in the 1980s. AI is used in many contexts so everyone has been involved with it in some way, including me.",3,3,2,2,3,1,1,4,"Difficult, because I have very little experience of computer games myself.","I understood the questions, but in my assessment I thought a human could make the same mistakes or draw the same conclusions as a machine.",2,4,4,4,2,1,3,3,4,Faster than the previous one. Obviously got to the finish line,This was easier when there was a reference point from the previous one.,4,1,1,1,5,4,1,5,1,"Easier, because the discussion showed that it follows a formula and does not know how to apply it. ","This was clearer to answer.,4,1,5,5,1,1,5,1,5,The AI seemed to be able to adapt to the changing environment.,I think I understood pretty well.,5,Re-checked the material 1 - 2 times,Important research if it is useful for the development of AI in general.
5.2.2023 klo 21.29.58,18 - 24,Female,5,I have played the game extensively,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,1,3,1,3,5,1,4,2,,,4,5,3,3,3,1,1,1,3,1,,,4,1,3,1,4,5,1,3,3,,,3,5,4,1,2,1,1,1,4,,,3,"Only watched the material once, did not return to it",
10.2.2023 klo 10.03.49,Over 50,Female,3,I have played the game a little bit,I have used a program that utilizes AI techniques (e.g. ChatGPT),"Super Mario Bros.: my kids (now adults) played it a lot, and I tried so hard to keep up. No chance. It was humiliatingly hard for me to play.
AI: I work in a science centre, where we have AI-themed exhibition. Working with that has gained me some experience, as has the Helsinki University AI-course, which I took.",4,4,5,1,1,3,3,5,Interesting questions,No.,4,5,5,4,1,1,4,2,5,"Easy, as AI was playing like a pro",No,5,4,4,5,1,1,4,3,5,"Interesting. Adding the conversation made a big difference, humanising the AI",No.,4,5,5,5,1,1,4,3,5,"Interesting again. It is always still surprising how well AI is responding on conversations. It makes me wary (and a bit suspicious) of the situation. When talking online with someone, is it really someone or just something. Does that matter if I don't recognise the difference and feel good afterwards?",No,4,Re-checked the material 1 - 2 times,"This was an interesting survey, which led me on rather basic questions about making a difference on AI and humans - is there a difference, and does it matter?
in the end it is scary to see how happy old people suffering from memory loss are, when they are given a furry robot or a smiling robot. I really hope it will not be me in the future."
10.2.2023 klo 10.52.47,41 - 50,Male,4,I have played the game a little bit,I have used a program that utilizes AI techniques (e.g. ChatGPT),"I know a lot about the theory behind neural networks. I also know about the different learning types and training models. My AI knowledge ""ends"" after first CNN models.",3,2,1,4,5,2,3,2,"I felt like that this is ""evolutionary model"" where AI has learned to play only via trial and error and repeats its mistakes despite what's seen on the screen. It probably doesn't understand what it is doing.",not at all. I think I know a little bit more about AI than I first stated?,5,4,3,1,2,1,5,4,2,"It felt like an optimised version of the previous video AI, and I considered that the goal is to get to the castle - but it missed a lot of other items in the playfield (kill all, collect all).",Nope.,5,2,3,3,4,5,4,4,3,"I bit more difficult as I was focusing more on the chat log than actual gameplay (which was pretty bad). So it is a bit more difficult to answer these questions as I need to find a compromise between the ""chat AI"" and ""gameplay AI"",,4,5,4,1,4,2,4,2,3,"Nice, because it didn't feel ""gringy"" :D It did ok.",,4,"Only watched the material once, did not return to it",Was the goal speed run Mario?
10.2.2023 klo 13.18.04,25 - 31,Other,5,I have played the game a little bit,I have used a program that utilizes AI techniques (e.g. ChatGPT),"I am more culturally familiar with the Mario franchise, and have not played it almost at all. Nevertheless I am very familiar with it via cultural influence. ",1,4,3,2,4,5,2,5,Fun -- I instantly applied a personality to the AI and humanized it. ,,4,5,5,3,1,1,5,2,5,"The video was sped up, so the nuances in character control were harder to perceive. Thus it was harder to humanize the AI playing the game. ",,4,1,4,4,4,5,5,5,2,4,,,2,4,4,4,2,1,4,2,5,,,3,Re-checked the material 1 - 2 times,
10.2.2023 klo 14.26.57,25 - 31,Male,5,I have played the game extensively,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",I am a casual gamer and I have grown up along with the evolution on gaming consoles. Super Mario Bros. was the first console game I ever played and have played it for countless of hours (but still never got through it).,2,2,4,2,4,3,3,4,"Easy and straightforward. Seeing how the AI performed, it gave me some sort of an idea behind the logics in its behaviour.","Yes, the ones with the questions about surprising me. The AI just did not surprise me with its behaviour, for I had no expectations.",5,5,4,4,1,2,4,2,4,"Quite easy, because it seemed to have a specific goal - to get through the level as fast as possible.",No.,4,1,2,4,4,5,4,2,4,"Different than previous times. This time it explained what it was doing oor ""feeling"" about the game.",No.,4,5,4,4,1,2,4,2,4,"Interesting, but quite similar to the phase where I could not see the messages.",No.,4,"Only watched the material once, did not return to it",The AI felt more human when there were messages and explanations about its behaviour.
10.2.2023 klo 15.26.48,25 - 31,Male,4,I have played the game a little bit,I have used a program that utilizes AI techniques (e.g. ChatGPT),,4,2,2,5,5,1,4,4,It was interesting. I was mildly frustrated to see the AI miss most of the ? boxe.,No.,5,5,4,5,2,1,4,2,5,"I was first frustrated to see the AI miss again many boxes and foes, but then remembered that there is this concept called a ""speed run"" and after that it seemed to me that the AI was aiming for maximum speed and its behaviour started to make much more sense after this realization.",No.,5,1,4,4,2,4,2,4,4,Humanlike chat from the chatbot made me interpet things differently. I assume I might be watching the same video as in the first part of the questionnaire though I am not sure.,No.,4,5,5,5,1,2,5,1,5,I thought the AI might have gone for speed run this time already before the start of the video. Now I think I might be more confidant in the fact that it was what the AI was aiming for and that it achieved its goal.,No.,5,Re-checked the material 1 - 2 times,It was a really interesting survey! I wish you all the best with your thesis!
10.2.2023 klo 22.22.29,18 - 24,Other,5,I have played the game extensively,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,1,1,1,5,4,1,5,1,"a bit hard, worried im rating wrong",no,4,5,5,1,1,1,5,2,2,surprised,no,5,1,3,4,3,3,1,2,3,"hard, confusing","the ai chat made it harder for me to understand what was happening because the ai referenced the same time and gave different answers to the users questions,3,5,5,2,1,1,3,1,5,,,2,Re-checked the material 1 - 2 times,
11.2.2023 klo 0.28.04,18 - 24,Female,5,I have played the game extensively,I have used a program that utilizes AI techniques (e.g. ChatGPT),,3,2,1,5,5,2,3,2,,,5,5,5,2,1,2,4,1,4,,,5,3,4,4,3,5,3,3,4,,,5,5,5,4,1,3,4,2,4,,,5,Re-checked the material 1 - 2 times,
11.2.2023 klo 0.31.44,18 - 24,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI techniques, I have read or watched about AI",I watched the Sethbling Mario AI video.,3,4,2,2,5,4,2,4,It was interesting.,"The AI did something negatively surprising, took a bit in answering.",4,5,4,2,2,3,4,1,4,"Tbh it was funny to see ""I could play better than the AI""",Nah,4,2,3,2,4,5,2,2,2,The chatbot vs ai was interesting.,It was kinda hard to tell that the chatbot wasn't playing the game. It seemed to spam accurate messages which made me double take.,4,5,4,1,2,3,2,4,4,the random chatbot messsages made the run feel goofier.,"not misunderstood, but the red chatbot message felt vague and as if it could apply anywhere to this perfect run.",4,Re-checked the material 1 - 2 times,"The screenshot of the messages in video was useful for not having to look back so many times, nice work.

The Chatbot AI seemed kinda weird, didn't really know what the blue bot was responding to. "
11.2.2023 klo 0.43.15,18 - 24,Male,5,I have played the game a little bit,I have read or watched about AI,,2,4,2,2,3,2,2,4,,,5,5,5,5,1,1,5,2,5,,,5,2,3,3,2,4,3,4,3,,,5,5,5,4,1,2,4,2,5,,,5,"Only watched the material once, did not return to it",
11.2.2023 klo 0.46.05,18 - 24,Other,4,I have played the game a little bit,I have used a program that utilizes AI techniques (e.g. ChatGPT),,2,2,2,4,5,4,2,3,,,5,5,5,4,2,2,5,4,4,4,,,5,1,4,2,2,5,5,3,4,,,5,5,3,2,2,5,4,2,4,,,5,Re-checked the material 1 - 2 times,
11.2.2023 klo 6.37.09,18 - 24,Male,3,I have played the game extensively,I have read or watched about AI,"Owned the game as a child, played and watched it be played for literal hours.

Regularly steak and consume articles and video essay's on how an AIs are preforming, where they've been and where they will likely go.",1,2,4,2,4,5,1,5,"Emotionally, it was endearing. The positive surprise was seeing the AI, in my option, fail. As I assigned the win condition before even watching the video as getting to the end of the lvl. It was like watching someone new to video games trying to do a Speedrun but being unsure if they wanted to commit. If I where to watch further videos about the same A.I. learning I would assign the win condition as 'Get father than the last run'.

Logically, it was surprising. I had an expectation the A.I. would behave in a point by point perfect manner. ",I wish there was more opportunity to provide *why* the choice was selected.,4,5,5,1,1,1,1,4,3,"It felt like what I expected when told ""Watch an A.I. play"", contrasting to the previous run.

I have confronted my feelings on ""overly mechanical"" AI, and am neutral to positive about it based on the context, but every time I *actively watch* an A.I. become more streamline I experience sadness at the ""loss"" of it's ""personality""

Perhaps I am influenced by media, personal struggles with self expression, or some other such in having such a inate response.",N/A,5,1,4,4,1,4,2,2,2,"Oddly enough, the blue messages felt less ""human"" than the red messages.",N/A,4,1,2,4,1,4,1,1,4,"I feel like the red text feels more ""human"" due to it not sounded like ""Mario"" who I've known for a long time is fictional ",,5,"Only watched the material once, did not return to it",I did not rewatch the videos due to wanting to preserve my original thoughts and feelings
11.2.2023 klo 6.54.26,18 - 24,Other,4,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",I'm a big Mario fan,1,2,1,3,5,4,5,3,"As if its only goal was to reach the end as quickly as possible, ignoring items and enemies unless hit by accident.",,5,5,4,1,2,3,5,1,5,,,,5,1,2,2,4,5,1,4,3,,,4,5,4,2,2,2,5,1,5,,,,5,"Only watched the material once, did not return to it",
11.2.2023 klo 19.43.32,18 - 24,Other,4,I barely know anything about Mario,I have read or watched about AI,"I am familiar with many video games but have no experience with any Mario games whatsoever. I have never interacted with AI firsthand, only heard of it and seen things it has produced. ",3,2,1,4,4,1,4,2,"I felt a little confused, mostly because of my lack of experience with the game itself. If I understood more about what the goal was, I might have been able to evaluate more clearly.","I don't know what the goal of Super Mario Bros is, so I don't really know if the AI ""reached its goal"". Because of the timer in the top right, it looks like you're supposed to reach the end of the track as fast as possible, but I can't tell if you're supposed to kill the enemies or avoid them. Also, I don't know what the blinking ""?"" boxes are for. This also carries over to ""I could play better than the AI"" - maybe if I learned how to play the game, I would, but as it stands I don't know if I would be better.",2,4,3,3,2,3,4,4,3,"My answers for this and the next question are pretty much the same as the ones on the past page. I don't really understand the goal of the game, so I'm not sure how to tell if the AI did a good job.","It did get to the end of the level as fast as possible, which seems intelligent? But I'm guessing there's more to the game than that, so it could probably do better.",2,1,2,2,4,4,2,5,2,"I'm interested now in why the AI makes the decisions it does, as it seems to be following a preconstructed path or program of what it should do. However, it seems to have made a silly mistake that it has not made before.","I can tell that falling into the hole is not one of the goals for Super Mario Bros, so I'm less uncertain of whether the AI reached its goal. But it still seems to not be able to react to its environment very effectively.",3,4,4,3,2,3,4,3,2,"I like knowing that there's a reason for its actions, even if it still doesn't seem to be interacting very well with its surroundings.","I'm still trying to figure out if it's supposed to be doing something with the boxes or the ""goombas"". That would probably influence my response to whether it is doing well, reaching its goals, behaving intelligently, etc.",3,Re-checked the material 3 - 4 times,"I think it would be nice to get a primer on the game Super Mario Bros at the beginning of the survey. I appreciate that I was probably not the target audience for this, but I think experience with Super Mario Bros might be recommended rather than just helpful. I liked having insight on how the AI communicated and justified its actions. I hope the rest of your research goes well!"
11.2.2023 klo 21.53.22,18 - 24,Other,5,I have played the game extensively,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,1,3,1,2,5,2,2,2,,"I wasn't sure what exactly was meant by the ai being ""dumb,"" or the difference between positively and negatively surprising.",4,5,3,2,2,3,2,4,3,"All of these questions are hard to answer honestly, as the simulation of the game isn't exactly the same as the original.",2,1,4,1,1,5,2,2,4,,,3,5,4,2,1,1,2,2,4,,,4,Re-checked the material 1 - 2 times,
11.2.2023 klo 22.20.22,18 - 24,Other,4,"I have seen/played other Mario games, but I have not seen or played Super Mario Bros.","I have used a program that utilizes AI techniques (e.g. ChatGPT), Currently receiving education in the field of computer science, have passingly speculated on how to create an AI",,3,4,2,2,2,2,4,5,"It felt a little bit like sticking my hand in a pot of slime to observe how it would react to my interference. It was somewhat surprisingly uncomfortable, but also quite fascinating","I'm not sure what would constitute as either positively or negatively surprising. I'm also not sure what the objective was that the AI was given, or what previous knowledge it had of the game, if any.",4,4,3,3,2,1,5,1,4,"Fun. It's enjoyable to watch a task be performed well, and all the more satisfying to watch it be completed quickly.",,4,4,4,2,2,4,4,3,4,"The AI is a lot cuter when it can print messages. Talking while playing a game is a very organic-feeling quality. The actual phrases felt very computer-generated, but they were somehow just as charming, if not more so, than if they had felt totally natural.",,4,4,3,4,2,1,4,1,5,"It was somehow less exciting than when there was no chatbot, but I think the chatbot (or chat function) still adds an interesting level of interaction with the AI.",,4,"Only watched the material once, did not return to it",hehe computer go brrrrrr <3
12.2.2023 klo 0.23.35,25 - 31,Other,5,I have played the game a little bit,I have read or watched about AI,"I've watched several videos of 2D Super Mario games being played by AI using neural networks.

I've played and watched others play many Mario games growing up. I have some understanding of AI, but no background in software, so my knowledge is limited.",2,3,4,4,3,4,1,5,"Strange, I guess. I feel like it was doing a lot of the things a brand new player would do, learning the controls in real time, learning how to avoid obstacles, and pausing to assess situations. But it also had certain behavior that felt mechanical, such as the time it took to understand stairs. A human observer may have a more intuitive idea of slope, and may place more value on exploration or getting power ups, which the AI wouldn't have any incentive to do or not do.","Yeah, it's hard to know whether the AI met its goal without knowing what kind of scoring system its software uses. I assume the goal was to get to the end of the level, which it didn't do, but it's sub-goal of learning to navigate the level was reached very well. I also don't know how to judge whether it was acting ""intelligently"" or in a human-like manner. I said yes to human-like, because it did have certain behaviors that seemed human-like, and it did seem to learn those behaviors through observation and practice. Intelligence is much harder to measure because defining it is difficult even for humans. The AI followed logical steps to learn a new skill, which is intelligence. But it took time to learn that skill and it didn't make certain logical steps to improve, such as keeping a running speed going. I think when people think of 'intelligent' machines they think of genius level software that does things better and faster than an above-average human. It wasn't that, but it was about a child's level learner. So it depends on how we're defining intelligent.",4,5,5,2,1,1,5,1,5,"Easy this time. This looks like a Tool Assisted Speedrun, with technical perfection at beating the level at maximum speed.",No.,5,1,4,4,5,4,1,5,5,"Similar to the first question. It seems like the AI was learning as it went, and this one was more timid, going backwards on landing before it went forward again. Having a little more insight was interesting. For a moment I considered it less intelligent because it was using hard-code patterns, but I realized it was still choosing between those patterns, so it was still reacting to it's environment.",,5,5,5,1,1,1,4,1,4,"Strange in a different way. I know logically that the chat function of the AI's explanation was written by a person from a dialogue tree, but AI using language still feels very human-like.",No.,5,Re-checked the material 1 - 2 times,GLHF
12.2.2023 klo 9.58.03,25 - 31,Other,5,I have played the game a little bit,I have done research involving AI,"I'm familiar with the basics of Super Mario Bros. but haven't personally played it in a long time. I do recall reading about an AI playing the game and finding some new exploits or glitches from about 4 years ago or so. I've studied some artificial intelligence, mostly focusing on artificial neural nets and genetic algorithms.",2,1,2,4,5,2,2,3,It felt a little difficult being confident in some of the answers as the test itself doesn't evaluate individual skills that the game would require but instead showed a single play through by the AI. It's unusual play style is unsurprising after seeing similar AI work done many times before.,,3,5,3,3,3,1,1,1,3,"While it would appear the AI had a flawless play and certainly beyond my own capabilities, I am wondering how much of this is instinct or intelligence. It is the same level and surely had a lot of experience in trial and error, more than a human would need to get to the same level I suspect.",,5,1,1,1,5,5,1,1,1,I felt more confident with my suspicions confirmed that the AI is operating without a robust system of reflection and reasoning.,,5,5,4,3,2,1,4,1,4,The text from the AI doesn't really convey understanding of the situation beyond instinct still though the chatbot mario accurately recognizes the reason based on context. But I am still left wondering how deep that understanding actually goes.,,4,Re-checked the material 1 - 2 times,"I think it's important to develop AI to work less as a black box and communicate and cite it's own reasoning more effectively, especially in an effort to make AI safer to use in other applications."
12.2.2023 klo 10.52.49,25 - 31,Female,5,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,4,4,1,2,5,1,2,5,"was succinct, and included all questions i can think of that were relevant, except maybe ""did the ai achieve its goal"" which is, from its perspective, i assume yes. it had the most success from its peers of moving right, but from our perspective no. it did not complete the level and died. It achieved its goal, but not ours.",the aforementioned first question ,5,5,5,4,1,1,5,1,5,easy this time,no,5,1,2,1,4,5,1,2,2,bit more muddy but these questions are making sense at this point,no,4,5,5,4,1,1,4,1,5,good and clear,no,5,"Only watched the material once, did not return to it","very interesting ai being developed here, artificial vtuber incoming?"
12.2.2023 klo 20.31.49,25 - 31,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have read or watched about AI",,2,2,4,4,2,2,4,4,,,3,5,5,1,1,1,1,1,1,,,5,2,1,2,5,5,2,4,2,,,4,5,5,3,1,1,4,1,4,,,4,"Only watched the material once, did not return to it",
13.2.2023 klo 15.02.01,25 - 31,Female,5,I have played the game a little bit,I have read or watched about AI,,1,3,2,2,5,1,1,2,,,5,5,4,3,1,3,3,1,2,,"Its pretty difficult to tell if the AI plays human-like when there are multiple ways to play a game. I also didn't think that the AI felt dumb, but I can't really tell if its smart either or whether it just played the stage enough times to know it.",4,1,2,2,1,5,1,1,4,,,4,5,4,3,1,4,3,1,4,,,4,"Only watched the material once, did not return to it","Its kinda hard to determine whether the AI playing Mario is smart or if it just maps out the stages its playing and then doing the bare minimum needed to succeed. In a way, this AI reminds of Minecraft Baritone pathfinder bot that quickly determines the fastest routes it can possibly take to reach whatever goal it is set to reach."
13.2.2023 klo 15.28.10,32 - 40,Female,3,I have played the game a little bit,I have read or watched about AI,,3,1,1,5,1,4,4,2,"It was kind of hard to evaluate the AI when I didn't get any information on how the AI was ""trained"" and what was it's strategy. For example, ChatGPT writes one word at a time and evaluates after each word what could be the next suitable word based on all the material it uses. Was this AI doing the same? Moving a little bit and the evaluating what the players usually do next? Did it actually react to the things happening in the game, or was it just reflecting what the previous players have done in this field? And what was the goal? To pass the level? To collect everything from the level (not likely)? ","Not really. Although the question ""The AI behaved in an intelligent manner"" would be better to phrase ""It felt like the AI behaved in an intelligent manner"", as we know that AI does what it is programmed to do, and that there's no intelligence, choices or will in it.",4,3,1,1,1,1,4,4,2,Easier than the first one.,,4,3,1,1,4,1,1,1,1,"I still don't know anything about the AI's programming, and what kind of a strategy it's following, so kinda hard.",,3,4,1,1,1,4,1,3,,,3,Re-checked the material 1 - 2 times,
14.2.2023 klo 15.12.40,32 - 40,Male,5,I have played the game extensively,"I have done research involving AI, I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI techniques, I have read or watched about AI",Not much familiarity,2,2,2,4,5,3,2,3,Quite ok,No,4,5,2,2,3,3,4,2,2,Quite ok,No,4,3,2,2,4,5,2,4,2,quite ok,no,5,4,2,2,3,4,4,3,2,quite ok,No,4,"Only watched the material once, did not return to it",
14.2.2023 klo 15.24.17,32 - 40,Male,5,I have played the game a little bit,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI techniques, I have read or watched about AI",,1,4,1,3,4,3,5,5,,,4,5,4,3,1,1,5,1,1,,,3,1,1,3,5,5,1,4,2,,,3,5,5,3,1,1,5,1,4,,,3,Re-checked the material 1 - 2 times,
14.2.2023 klo 16.25.27,32 - 40,Female,3,I have seen videos of someone else playing the game,"I have used a program that utilizes AI techniques (e.g. ChatGPT), I have created or helped with creating a program that utilizes AI techniques, I have read or watched about AI","I have worked with localisation and quality control of different Natural Language Processing solutions.",1,3,2,4,5,1,4,4,,,5,5,5,2,1,1,4,1,4,,,4,1,2,3,3,4,3,4,4,,,4,5,5,3,2,1,3,1,2,,,4,Re-checked the material 1 - 2 times,
19.2.2023 klo 1.29.32,25 - 31,Female,2,I have played the game a little bit,I have read or watched about AI,,4,4,3,1,3,4,1,3,It makes me think about how it works in video games,I am not professional with AI so judging the video is hard,3,2,3,1,1,5,1,5,1,It is surprising for me now after the 1st video,no,5,3,3,1,3,5,2,1,5,It is hard to evaluate,no,5,5,5,2,3,3,5,3,5,It is very confusing now,no,5,Re-checked the material 1 - 2 times,
1.3.2023 klo 16.33.45,41 - 50,Male,3,I have played the game a little bit,I have created or helped with creating a program that utilizes AI techniques,"Never had a nintendo myself, but have played Mario Bros occasionally, and a lot of other platformers. In the AI front, I've done some solution development including creation of custom training visualisations and building pipelines for specific tasks.",4,3,1,2,5,4,3,3,It is not quite straighforward to evaluate,Maybe the questions that explicitly ask about whether the AI felt dumb is a bit in the antropomorphist category.,4,5,5,1,1,1,5,3,5,"When the goal of the AI is clearly visible in the process, it feels easier to judge it. In the previous example where the model had been trained for not so clearly single goal, it was much harder to tell how it actually did.",,5,3,3,2,4,5,3,3,4,"This was more fun than the ones with only playback. The model shouting out what it does was a fun element, but not necessarily very clever. In the aftermath the model did not feel like being too good in reflection why it would have done what it did, the explanation being just about choosing a behaviour x based on pattern y.",,4,5,4,1,3,3,3,3,4,It starts to feel like something that I don't want to carry on doing :D,-4,Re-checked the material 1 - 2 times,"This was an interesting survey, and quite a bit of fun to watch AI try to perform a leisure task."
21.3.2023 klo 11.52.44,25 - 31,Male,5,I have played the game extensively,I have used a program that utilizes AI techniques (e.g. ChatGPT),,1,2,3,4,5,3,3,2,A bit difficult,,4,4,4,3,2,3,5,1,4,,,5,1,4,4,2,4,3,2,5,,,5,5,4,3,1,4,4,2,5,,,4,Re-checked the material 1 - 2 times,