Uncanny Logic

A theoretical extension of the Uncanny Valley of the Mind hypothesis influenced by the Perception of Intelligence in Black Box Artificial Intelligence

Milan Mrdenovic

Candidate No. 5

Master's Thesis in Screen Cultures

Department of Media and Communication
University of Oslo

Fall 2021

Uncanny Logic

A theoretical extension of the Uncanny Valley of the Mind hypothesis influenced by the perception of intelligence in Black Box Artificial Intelligence

Milan Mrdenovic

Supervisor: Taina Carola Andrea Bucher (Fall 2020-Spring 2021)

Abstract

In this thesis, I posit a conception of a specific subset of the Uncanny Valley of the Mind sensation that I call Uncanny Logic. A notion that was inspired by a passage in Adam Greenfield's book *Radical Technologies: The Design of Everyday Life* detailing the experiences of discomfort by the spectators who witnessed the Go matches between Google DeepMind's AlphaGo and Lee Sedol in 2016. This is done through a critical discussion of the current day technological reality of artificial intelligence as a field and its conflicts with the perception of artificial intelligence as an object of the uncanny. Via these discussions, I attempt to codify a tentative definition of the concept through analytic induction, which is thereby filtered through four alternating case studies. Two of these case studies are real-life events, and two are taken from fictional media so as to further illustrate the divide between current-day technological realities and perception with the cultural imaginaries of intelligent machines. Their insights are utilized to amend the final definition of the concept, which I claim will be of importance for understanding such complex feelings of unease in the future as our societies become further saturated with artificial intelligence-based technologies.

Acknowledgments

Many people have helped me get through the mentally and emotionally laborious process of writing this thesis. Firstly, I would like to thank my supervisor Taina Bucher. Your advice was instrumental in shaping the vision for this thesis, and for that, I am grateful. Another mentor I wish to thank is Zuzanna Zygadlo; your support and perspective helped me strike a balance between what I needed to do and what I wanted to do. I am grateful for the support of my friends Arjeta, Ana, Milica, Nikola, Basak, and Lali, whose late-night chats and pep talks kept me from losing my way to my goal. Thank you to my love Filip for your constant love and support, as well as your faith in my ability to handle this challenging undertaking. And last but not least, thank you to my family. Mom, Dad, and Brother, without your support, I wouldn't even be here to undertake this task in the first place. I love you all.

Contents

1 Introduction

"But there was something almost numinous about AlphaGo's play, an uncanny quality that caused at least one expert observer of its games against Lee to feel "physically unwell"."

- Adam Greenfield, Radical Technologies: The Design of Everyday Life

Through this sentence, Adam Greenfield describes the seemingly irrational manner in which DeepMind Technologies artificial intelligence system AlphaGo played against and decimated (4 – 1) the world champion Go player Lee Sedol in 2016 (Greenfield, 2018, p. 238). The spectators of these events were plagued by great discomfort over the capabilities of the artificial intelligence system since the game was considered to be unbeatable for an AI at a professional level (Kohs, 2017).

This description of the event ignited the search for this notion and the creation of this research project. A sensation reminiscent of the *Uncanny Valley* hypothesis, albeit seemingly originating from a very different source, not the appearance of the object in question, yet rather its incredible *"mind."* What could lay behind this sensation of encountering a logic or rationale that would baffle the human mind and what it could mean for our future? As our technology progresses, through building more elaborate systems and applying them in ever more common circumstances, could we encounter this feeling more and more? Was this an isolated case mediated by its particularities as a seminal event in artificial intelligence development? Or could this nebulous sensation be found in other places, perhaps even within our cultural imaginaries of intelligent machines (Elish & boyd, 2018, p. 66)?

Yet, there was one glaring issue; the notion was undefined. The sensations the participants and spectators of these events espoused were all too vague and based on a general unease or existential malaise over what they had witnessed (Herschberger, 2021; Kohs, 2017). Therefore, the notion required definition; it needed limitations and concrete description of its elements if it were to be explored at all; as such, it had to become a concept, one that I call *Uncanny Logic*.

The goal of this research project became the development of this concept so that its main research question could be answered – "*How could the idea of Uncanny Logic affect interactions with and perception of highly advanced AI*?". If this type of feeling could one day become an endemic part of our lives, our vocabulary would need to expand in order for discussions surrounding these types of events to be a possibility. Through a pair of literature

reviews, elements of importance for the construction of the concept are divulged. Each literature review is focused on a specific element of the discussion regarding AI; one tackles its technological reality and the other the issues of perceptions of AI as uncanny.

A close contender for the notion existed within the *Uncanny Valley of the Mind* theory, wherein the discomfort related to machines is connected to perceiving a mind in the machine (more specifically, either its emotional or agency-related sentience) (Gray & Wegner, 2012). However, while it worked as a great base for a concept of the notion related to the previously specified event, it lacked nuance for some of its most important elements. For this reason, the concept I was developing would entrench its base inside the theory and function as a theoretical extension of it by accounting for the specifics of the cases in question.

Two characteristics of advanced artificial intelligence systems based on artificial neural networks (ANN) became increasingly important for developing the concept of Uncanny Logic. Firstly, the complexity of the systems themselves being sufficiently great to exhibit the emergence of new properties not normally found within parts of the system. And secondly, the opaque nature of these systems both from the side of lack of familiarity with the technology in the general population and the incompatibility of ANN decision making when seen from the perspective of human cognition.

This firmly stations the concept within a conflict of perception of the artificial intelligence systems in question and their technological reality. This distinction is important since issues rooted in perception may someday become issues of communication. We may not have sentient AI today, but if they ever became a reality, we would be faced with individuals or groups that have a completely different frame of thought, whose decisions or motives are unknowable to a human. Seeing as how miscommunication and distrust have led to conflicts through human history, it wouldn't be hard to imagine (as science-fiction has already explored in-depth) what would happen if we were to contend with a new *"other"* group.

The development of the concept was based on an Analytical Induction approach where a tentative definition is used to describe the concept in question before it is compared to the representations within the case studies. The insights from the case studies shape and amend the conception of the notion before the final definition is established.

To illustrate the element of possible future conflict, real-life current-day examples of AI and their perception by the public are contrasted with examples from fictional media. This is done through the analysis of four cumulative case studies that serve as the basis for an

analytical induction approach to developing the concept. The selection of case studies includes two real-life case studies (the aforementioned AlphaGo and the OpenAI Five) and two case studies based on works of fiction (The Puppet Master from *"Ghost in the Shell (1995)"* and HAL 9000 from *"2001: A Space Odyssey (1968)"*). The studies alternate between real-life and fictional cases to elucidate the contrasts between them and improve the flow of insights from one to the other. Through a mixture of textual and critical discourse analysis, parallels are drawn between the events, representations of the AI, and the discourse surrounding them. Ultimately culminating in the defining of this all too nebulous notion through the lens of these four instances.

2 Methodology

The Methodology chapter specifies the methodologies and techniques used for this research project as well as its general structure. Due to the interdisciplinary nature and theoretical focus of this thesis, a varied selection of methods conducive for this type of research was selected. As the focal point of the thesis is the development of a concept that could be grafted on as a theoretical extension of a prominent theory within several disciplines, the research methods, while of a qualitative nature should result in a concept that can travel well between varied disciplines. The choice of qualitative methods over quantitative is mainly due to their propensity for better answering humanistic questions of a how or why nature (M. N. Marshall, 1996). As such, they will be more useful for explaining the complex interplay between the elements of the discussion. However, the argument for the interdisciplinarity of this text should not be taken as an abdication to a chaotic self-indulgent approach in the completion of the work (Bal, 2002). For this reason, the concept will be developed through a framework of analytic induction coupled with a set of comparative case studies based on the analysis of media texts either within fictional media (science-fiction film) or documentaries focused on real-life events; this is meant to encapsulate the development of the idea to a stricter structure.

2.1 Concept Development

The goal of this thesis is to create a concept that extends the notion of the Uncanny Valley of the Mind into a specific domain of experiencing unease about the intelligence of the artificial object in question. Thereafter, working on answering the main research question of this text – *How could the idea of Uncanny Logic affect interactions with and perception of highly advanced AI*? However, in order to achieve this goal, I must first specify what is meant by the word concept within this context.

According to Mieke Bal, concepts are highly interpretative *"tools of intersubjectivity"* that serve as abstractions of objects that allow for complex discussion through shared language; thereby, in a certain sense, they represent small-scale theories (Bal, 2009, p. 19). They are always composed of multiple components (often other concepts) and are created with the purpose of elucidating problems that are considered to be misunderstood or incompletely understood (Deleuze & Guattari, 1994, p. 17). To serve this purpose, they need to be clearly defined and explicit in their meaning while at the same time offering a great

dose of flexibility required for them to be applicable in various situations (Bal, 2009, p. 19). If the concepts are too rigid, their applicability is limited, whereas if they are too fluid, they lose the cohesiveness required for their status as a concept. Thereby they seemingly require a balance between two contradictory states. They are never created from a void; they always have a history and are often dependent on interactions with other concepts to give them meaning, which occurs on intersections between concepts (Deleuze & Guattari, 1994, p. 18). The ultimate goal of a concept is to demarcate an idea and focus interest within the target of study (Bal, 2002, p. 31).

An important distinction exists between concepts and functions. Functions represent the functional properties of things or objects (Deleuze & Guattari, 1994, pp. 117-119); in the case of this text, we could say that an artificial neural network has the function of solving equations to create algorithmic predictions. Yet, we cannot say that an AI has the function of being uncanny; it is not a proposition of the state of affairs or an intrinsic functionality of the object in question. Instead, it is a complex subjective experience elucidated far better by concepts and the discussion they invoke than functional analysis. This boundary between functions and concepts will be illustrated throughout the work, as examples of both are encountered.

The research question posited by this thesis thereby seeks a concept flexible enough for it to be capable of traveling between different disciplines. Concepts travel between different disciplines and audiences, thereby carrying with them differing nuances in meaning, scope, or perceived value (Bal, 2002, pp. 28-29). By traveling, they are also exposed to different kinds of argumentative and methodological scrutiny, any of which could damage or embolden their status within said research community (Bal, 2002, p. 29). Yet, Mieke Bal states that from the stance of normative epistemology – *"Concepts are legitimate as long as they avoid the status of 'mere metaphor' or ideology."(Bal, 2002, p. 29)* even if it is exposed to different types of scrutiny, their fundamental importance is their capability to facilitate debate. As such, individuals must decide on and justify their meaning and perceptions of said concept, for concepts enable – *"both a description and experimentation with the phenomena, which in turn allow for actual intervention, a new concept founds an object consisting of clearly defined categories"(Bal, 2002, p. 33).* Concepts that straddle the boundary between different disciplines (such as the Uncanny Valley) therefore must exhibit a great level of flexibility, and it is the responsibility of the researcher to correctly delineate and specify their interpretation.

The concept of Uncanny Logic, which is a fundamental element of the research question of this text, is dependent on several disciplines and concepts. These elements will be detailed within the literature reviews wherein the technological reality of artificial intelligence is understood from a computer and system science perspective while the societal perception of AI (in particular notions of uncanniness) is entrenched within the research of the humanities and social sciences. As such, a method that can synthesize the approaches of all three is the main tool that will be used for the development of the concept itself, namely – *Analytic Induction*.

2.2 Analytic Induction

Analytic Induction is a research method used for collecting, organizing, and analyzing data with the formal objective of creating a supposedly causal explanation for a phenomenon; this is achieved through continuous analysis of a sample of cases while progressively redefining the phenomenon that is being described as new insights are gleaned or as contradictions are encountered (Katz, 2001, p. 480). Its purpose is to seek out the particularities of any specific event or phenomenon (Judith Preissle, in Given, 2008, pp. 15-16). It was first posited by Florian Znaniecki in his book *The Method of Sociology* (1934) as a purportedly more scientific method for researching sociology inspired by the natural sciences, with the ultimate goal of establishing a deterministic causal explanation of the phenomenon (Znaniecki, 1934). He held the view that a phenomenon of interest does not need to be defined in advance of research; rather, it is to be unearthed through the process of study itself (Hammersley, 1989, pp. 161-163). However, over time with a healthy level of constructive critique, the approach shed its goal of yielding deterministic universal results, rather becoming a useful tool for creating notions and solid definitions (Katz, 2001; Ratcliff, 1994). This is due to the focus on inductive rather than deductive reasoning; the method does not permit such deterministic results, only general probabilistic conclusions inferred from specific observations (Ball & Thompson, 2018, p. 167).

A summarized version of the framework was detailed in six steps by Znaniecki's student Donald R. Cressey as follows:

"1) a phenomenon is defined in a tentative manner, 2) a hypothesis is developed about it, 3) a single instance is considered to determine if the hypothesis is confirmed, 4) if the hypothesis fails to be confirmed either the phenomenon is redefined or the hypothesis is revised so as to include the instance examined, 5) additional cases are examined and, if the

new hypothesis is repeatedly confirmed, some degree of certainty about the hypothesis results, and 6) each negative case requires that the hypothesis be reformulated until there are no exceptions."(Cressey, 1953 as cited by; Ratcliff, 1994). This explanation is a useful summary; however, it must be further deliberated upon for the purposes of this text.

Within the confines of Analytic Induction, a term need not be defined before the onset of research, as a definition is to be considered as a testable hypothesis; this allows for the concept to be altered during the research process in order to better represent the event or concept in question (Ratcliff, 1994, pp. 3-4). In a sense, the idea evolves alongside the research process and may need an amendment if it encounters unexpected contradictions or insights that question the originally proposed notion. As specified before, at first, a singular case is chosen to review and document the common elements and explanations for the phenomenon in question; once they are identified, the idea is to be contrasted against other instances of the supposed phenomenon (Hammersley, 1989; Katz, 2001). An important element of developing the concept is seeking out instances that challenge the initial case in some way; these are generally called negative cases, as they may either delimit or expand the scope of the theory (Judith Preissle, in Given, 2008, pp. 15-16). According to Katz – *"The logic of proof in AI relies solely on the richness or variety of the cases that have been shown to be consistent with the final explanation"*(Katz, 2001); as such, the quantity of cases is less important than their salience for exploring the notion in question. The previous goal of the method resulting in causal explanations for phenomena was predicated on the idea that Analytic Induction could be a tool of prediction (Znaniecki, 1934). However, critics of the original view of Analytic Induction like Katz have asserted that the more appropriate and influential purpose is not prediction, yet rather *"retroduction,"* the idea that in a retrospective analysis of a case if the phenomenon is observed in an event, specific requirements would have occurred prior to that (Katz, 2001, p. 483).

For the purposes of this text, Analytic Induction will be used as a general approach to exploring the idea of Uncanny Logic. The search for this notion was inspired by Adam Greenfield's passage about AlphaGo's manner of play - "*But there was something almost numinous about AlphaGo's play, an uncanny quality that caused at least one expert observer of its games against Lee to feel "physically unwell"."(Greenfield, 2018, p. 238)* as I did not possess enough knowledge about the topics at hand, it would not have been prudent to create a working definition without first heavily delving into the subject matter. For this reason, the first two chapters of the thesis are literature reviews that detail and discuss different elements and viewpoints in the discussion of both Artificial Intelligence and the Uncanny Valley.

Thereafter, the insights from both are synthesized into a tentative definition and explanation of the concept of Uncanny Logic. The notion is then reviewed within the context of the case that originally inspired Adam Greenfield's statement, that of DeepMind's AlphaGo playing against the world champion Go player Lee Sedol. Following the completion of the first case study, the definition is exposed to three other case studies, and their insights are applied to the concept in a sequential manner as this is considered to be the best approach in handling analytic induction (Katz, 2001). Once all of the case studies have been completed, the final definition altered by the influence of the collective case studies is presented and discussed alongside any closing remarks that are considered of importance for the text.

While this thesis is based on an analytic induction approach, I must also account for the possible discrepancies in my results that are born of this decision; there could have been a better way to explore this topic through, for example, a grounded theory approach. Perhaps, through a cyclical data collection/analysis process engaging with the individuals who were present for the events analyzed in this thesis (Kathy Charmaz and Antony Bryant. In Given, 2008, pp. 374-376). This could have yielded a differently nuanced view of Uncanny Logic. However, I found analytic induction to be more conducive for the type of thesis I had a goal of writing.

As case studies are of fundamental importance for this type of research, the reasoning and structure behind the approach taken within this thesis must be explained in more detail.

2.3 Case Studies

Case studies represent an approach where several instances of an event or phenomenon are analyzed in-depth, making them quite suitable for delineating ideas and building a more comprehensive understanding of a topic at hand (Joachim K. Blatten. In Given, 2008, pp. 68-71). As a research method, case studies suffer from several limitations when facing empirical evidence. They are difficult to replicate, are influenced by the subjectivity of the researcher, and can lack structured scientific rigor as they are highly interpretative. However, this very property makes them very useful for qualitative research, especially when analyzing media texts. This, however, does not mean that they lack any implications of causality when targeted at real-life occurrences; rather, they place more importance on providing a nuanced explanation of the event or subject/object in question than quantifying the elements of its causation (Joachim K. Blatten. In Given, 2008, pp. 68-71).

The following case studies will contain a mixture of:

- Textual analysis in relation to media texts such as films or documentaries depicting the events or representations which are being analyzed.
- General overviews of the technological basis of the AI in question (for the non-fiction case studies).
- Analysis of the representations through the lens of Uncanny Logic.
- Analysis of the social impact on the cultural zeitgeist related to these events or representations.

The goal of these case studies is to assist in answering the main research question of this thesis – "*How could the idea of Uncanny Logic affect interactions with and perception of highly advanced AI*?" by evaluating the definition of the concept against them. They are practically speaking a requirement of the analytic induction approach as, without a sample of cases, the method falls apart. This text utilizes four case studies, two of which are real-life cases of advanced artificial intelligence systems and the events surrounding them, while the other two are fictional representations of artificial intelligence found within film. The reasoning behind this is that exploring an idea as nebulous at first as that of Uncanny Logic requires the consideration of both the current-day reality of artificial intelligence technology as well as imaginary representations that echo human expectations of such technologies. As such, these case studies are comparative and created from a constructivist viewpoint, as I cannot attest to any universal generalization on part of this topic due to the complex interplay between the empirical evidence related to what artificial intelligence is and the way it is perceived or presented. Thereby the goal of these cases is bridging *"the gap between concrete observations and abstract meanings using interpretative techniques"*(Joachim K. Blatten. In Given, 2008, p. 69). The case studies alternate between real-life and fictional to provide balanced interaction between the examples, each of which feeds its insights into the other, allowing the concept to evolve on multiple fronts. An important caveat to mention is that the specific order of case studies influences the way insights are gleaned and framed, thereby shuffling the order might yield other insights not found within this text. The original case study of *AlphaGo* was directly inspired by the statement by Adam Greenfield as it seemed like the best starting point for searching for this notion itself. The other cases were chosen progressively throughout the research process due to their suitability for challenging the concept and providing it with the elements required for its further progression (Katz,

2001; M. N. Marshall, 1996). The other cases consist of the character *Puppet Master* from the film *Ghost in the Shell (1995)*, the artificial intelligence Dota 2 playing team *OpenAI Five* developed by OpenAI, and the character *HAL 9000* from *2001: A Space Odyssey (1968)* in that particular order.

The case studies follow two formats, one for each type of case:

The real-life case studies consist of:

1) *An introduction describing the general event that took place.*

This section of the case study is meant to ease the reader into pre-emptively understanding the focus and importance of the event in question, serving as the prelude for better understanding the complexities of the subject matter.

2) *A basic explanation of the game in question that the AI is taught to play.*

An explanation of the games in question is required to alleviate a form of opacity known as *Technological Illiteracy* (which will be further explained throughout this text) that relates to a lack of knowledge or understanding in a particular subject or field (Carabantes, 2020, pp. 312-313). Contrary to its name, which implies purely technological knowledge, it can also denote a lack of knowledge in, for instance, systematized rulesets of games. By providing this basic introduction, a reader should be able to follow the discussion of the events in question.

3) *A recapitulation of the development of the AI, a shorthand review of its original research paper, and the analysis of the technology.*

Understanding the development of the artificial intelligence system in question, as well as its real capabilities and characteristics, is of fundamental importance for understanding the contrasts between the reality of said system and the way it is perceived by the people who encounter or engage with them. The original research papers for each of the systems in question provide framed insights that contribute to this discussion both on the technological reality of the objects in question as well as the way the developers frame their creations in public discourse. *(See Critical Discourse Analysis sub-chapter)*

4) *Textual Analysis of the event through a documentary detailing the events.*

The selection of documentary films as points of reference for analysis allows for a limited discussion that enjoys some reproducibility. As these media texts are already released at the time of writing this text, they are more stationary examples of these events in comparison to interviews, blog posts, news reports, and the like. Another researcher may reliably return to

these examples and glean their own insights on the same sample used here without it changing over the course of time. *(See Textual Analysis sub-chapter)*

5) *Discussion of the event and its relation to Uncanny Logic.*

The case studies end with a recapitulation of important insights from the case study that build upon the notion of Uncanny Logic. In later studies, reflections upon the previous studies are detailed as they relate to the current example; this is meant to provide parallels of interest for further developing the concept itself.

<u>The fictional case studies consist of:</u>

1) *An introduction to the film and character of study.*

This section of the case study is meant to serve as a prelude to understanding the scope of the case study by specifying the film in question, its importance, and the focal character of the film being analyzed.

2) *A short summary of the plot of the film in question.*

While by no means an exhaustive summary, these sections are meant to introduce the reader to the plot of the film so that a general understanding may be established. The reader is, however, advised to experience these films themselves as this can yield a deeper understanding of this text and its subject matter if experienced firsthand. Nevertheless, a reader will obtain the necessary background information for following the case study from this framed account of the film.

3) *Textual analysis of the character in question.*

Much akin to the choice of documentary films for real-life case studies, the choice of films for fictional case studies should yield a static object of inquiry that is easily accessible to other researchers. However, in the fictional case studies, the focal point isn't the event itself yet rather the representation of the character in question within the film and its wider subtextual implications. By focusing on the character rather than the film itself, all elements that are unnecessary for the purposes of the topic of AI can be disregarded, such as the themes of transcending humanity in *2001: A Space Odyssey* under the influence of extraterrestrial technology. *(See Textual Analysis sub-chapter)*

4) *Discussion of the representation and its relation to Uncanny Logic.*

The case studies end with a similar recapitulation as the non-fiction cases, focusing on insights from the case study that build upon or change the notion of Uncanny Logic as well as how they influence the real-life cases when observed from a far-future science-fiction perspective.

All of the consecutive case studies also influence one another with their insights and the arguments made between them. The main goal of the case studies is not to provide step-by-step explanations of the causality of the phenomenon in question, nor could it be a possibility; rather, the intention is to utilize a body of empirical evidence alongside careful interpretive inquiry to construct a compelling argument for explaining the phenomenon in question (Joachim K. Blatten. In Given, 2008, pp. 68-71). This should result in a concrete definition of the notion in question that may be used to denote a particular conception of this type of phenomenon.

2.4 Textual Analysis

Textual analysis refers to the close interpretation of the content or meaning of a text with the goal of gleaning deeper insights of value for understanding the text in question (Saron Lockyer. In Given, 2008, pp. 865-866; McKee, 2003). The aim is not to claim a definitive interpretation; rather, it is about identifying possible or likely ones through analyzing of the cultural, social, contextual, or other cues observed within the text (Saron Lockyer. In Given, 2008, pp. 865-866). While the name textual analysis implies a focus on literary or written media, within the context of textual analysis, the term can also indicate audio-visual material such as film as well (Jennifer Morey Hawkins. In Allen, 2017, pp. 1754-1756). Within the confines of this text, this method is applied to the content of the media texts in question; within real-life cases, the focus is the documented and framed presentation of the event, and in the fictional cases, the representations of the characters in question. With close scrutiny the text can provide substantial information on the subtleties of the topic or object in question that can further the discussion and development of the argument being made (Saron Lockyer. In Given, 2008, pp. 865-866). When this method is applied in conjunction with critical discourse analysis it allows for a close reading of not only the contents of the media text, yet also a critical reflection upon the way these events or characters are presented to the public at large.

2.5 Critical Discourse Analysis

Critical Discourse Analysis (CDA) is an approach to critically examining the use of language that originated within linguistics yet branched out throughout the social sciences (Csilla Weninger. In Given, 2008, pp. 145-147). The object of scrutiny is the discourse itself, which Csilla Weninger defines as – *"generally understood to refer to any instance of signification, or meaning-making, whether through oral or written language or nonverbal means."* (Csilla Weninger. In Given, 2008, p. 145) or in other terms the use of language for the propagation or entrenchment of power relations. The critical analysis of the discourse aims to uncover and critique the subtleties of language that reinforce systems of power. Critical discourse analysis is mainly used within this text to frame and challenge the discussions over the reality of artificial intelligence and the conflicting cues given by different stakeholders from developers to governmental bodies and even researchers. As well as the views of machines as the ultimate other attempting to reinforce human supremacy over all other forms of life, be they animals or as of now non-existent conscious machines, all through the lens of the idea of the uncanny (a concept which has a history already weighted with unfair power dynamics). My goal is to challenge the societal perceptions of AI fostered both by those who would fear-monger against these technologies and those who would exalt their possibilities as impending reality.

2.6 Literature Review and Theory

A literature review serves to present and build upon the plethora of previous research in order to illustrate the debate within a field over the object of study in question (Richard Race. In Given, 2008, pp. 487-489). It delineates a body of work necessary to understand the topic so that a researcher may utilize this knowledge to build a case for their own research while at the same time providing a critical interpretation of said literature (Blaxter, Hughes, & Tight, 2006, pp. 122-123).

For the purposes of this thesis, the two literature reviews *"Artificial Intelligence – Theoretical Background"* and *"The Uncanny Valley"* serve simultaneously as presentations of previous work on the subjects, theoretical starting off points for the idea of Uncanny Logic and a critical discussion on the discourse surrounding artificial intelligence from varying perspectives. The former details the necessary information for understanding the technological reality of artificial intelligence today, shortly recapitulating the history of the

field as well as some of the larger challenges the technology faces. The latter focuses on the debate over the contentious idea of the Uncanny Valley, its different theories and criticism, as well as the perceptions of machines as the other through the lens of the uncanny. They provide the basis for a reader to follow the rest of the thesis as their insights synthesize to develop the concept of Uncanny Logic. Without them, there is nothing to ground the theory to any semblance of established knowledge, which would make it purely speculative and thereby invalid for any type of credible academic discussion.

The decision to have two distinct perspectives showcased is due to my view that clearly delineating the boundary between them enhances the richness of the discussion at hand, mainly because it firmly establishes the parallels and splits between them. Allowing not only for the discussion of them in isolation yet also the ways in which they interact or even the perceived boundary itself. Therefore, these segments of the thesis are the backbone of this research project.

The structure of the Artificial Intelligence literature review:

1) *What is AI?*

At the beginning of this chapter, I recapitulate a short history of the field of Artificial Intelligence research and development, as well as clarifying terminology and categorizations of AI in order to give readers the base knowledge needed to understand the discourse around the field of AI. Discrepancies found in the usage of certain terms are discussed since some individuals may not be aware of the nuance between terms like Strong/Weak AI and Artificial General Intelligence/Narrow Artificial Intelligence.

2) *Forms of Machine Learning*

This subsection is meant to provide a general understanding of the field of Machine Learning. Subjects such as Supervised, Unsupervised, and Reinforcement learning are discussed with examples to provide context for the development of Artificial Neural Networks. These AI systems are inspired by the functioning of the human brain, and they are an important element in the rise of Deep Learning (DL).

3) *Emergent Properties of Artificial Intelligence*

One of the fundamental elements of the concept of Uncanny Logic is the idea of Emergence - the capability of sufficiently complex systems to generate properties that are not a part of their base elements. Within the context of AI, this gives rise to a common issue known as Black Box AI, when a system becomes nigh impossible to understand even for

human experts. This subchapter also references and reflects upon a fascinating idea about emergence and natural artificial intelligence.

4) *Explainable AI (XAI)*

The occurrence of Black Box AI ignited the search for a way to reintroduce explainability into these complex systems since a lack of transparency in decision-making makes them highly dangerous tools to utilize. This chapter shortly recapitulates the history, discourse, and necessity for explainable AI due to its effects on public trust and danger to human safety.

The structure of the Uncanny Valley literature review:

1) *The Uncanny Valley*

The Uncanny Valley is one of the most influential yet contentious theories in the field of Robotics. This chapter opens the discourse on the Uncanny Valley by revisiting its origins, the original conception by Masahiro Mori as well as the refinement of his theories by other researchers into more useful conceptions.

2) *Criticism and Empirical Studies*

Understanding the criticism of the original idea of the Uncanny Valley and how it led to the development of new conceptions is of fundamental importance for the development of the concept of Uncanny Logic. A critical review of the search for empirical evidence for the sensation results in useful insights that delimit the possible scope of the idea, which is vitally important for determining the scope of Uncanny Logic.

3) *What is the Uncanny?*

Within his sub-chapter, I attempt to account for the contentious history and philosophical meaning of the word Uncanny and its powerful implications in the discussions over otherness. An argument is made for tying the notion of machine otherness and thereby their “uncanniness” to fears of challenging human sanctity or supremacy.

4) *Uncanny Valley of the Mind*

This conception of the Uncanny Valley excises the discussion over the Uncanny Valley from the field of aesthetics and into the perception of mind in the machine on the basis

of perceiving either emotion or agency. It becomes the basis for the concept of Uncanny Logic as the theoretical groundwork for a notion of perceiving intelligence in a machine.

Thereafter the literature reviews are synthesized into a chapter on the concept of Uncanny Logic, detailing its necessary elements from both sides, coupled with a tentative definition.

2.7 Ethics

Due to the nature of this thesis, no personal information was collected or stored; the basis of the case studies are published for-profit media texts in the forms of films (NSD, 2021); thereby, no application for personal data collection was submitted to NSD.

I declare that there were no competing interests in the creation of this work, no financial or non-financial incentives were received for or during the production of this text that may influence its findings or impartiality.

This thesis, however, must account for the bias of the author, as there is no possibility for truly disinterested and unbiased authorship to exist. I've done my best to objectively present the reality of the events and technologies in question while at the same time providing subjective commentary.

3 Artificial Intelligence – Theoretical Background

As a species, we have seemingly always been obsessed with the concept of creating life, both as a means of reflecting upon our own tumultuous past with wondering how we came to be and with our wish to meet another that could parry our intelligence. Our myths and legends are filled with skilled craftsmen bringing life to that which is lifeless. From the story of the Golem of Prague to the Automotones of Hephaestus in Greek Mythology, across cultures and continents, we find examples of this narrative.

In the modern-day, we can observe a culmination of this ancient drive in our attempts to endow our computers with the vague notion of intelligence, thereby creating the field of *Artificial Intelligence* research and development. The term Artificial Intelligence itself was first coined by John McCarthy in 1956. at a workshop hosted by DARPA (Heffernan, 2020, p. 93), the term mired the field it denoted in discussions over the nature of what is human and what is a machine as well as what the limitations of the concept could be. This led McCarthy to regret coining the term as such; instead, he believed that the more accurate terminology would have been *"Computational Intelligence"* – due to the notion of Computation as *"a calculation involving numbers or quantities"*(Heffernan, 2020, p. 93) which would firmly ground the concept within the field of algorithmics and mathematics without entrenching it in philosophical discussions about the nature of intelligence or being.

In the year 1950, Alan Turing released his famous *"Computing Machinery and Intelligence"* paper that attempted to develop the idea of an intelligent machine. For that purpose, he devised the *"Turing Test"* also known as *"The Imitation Game"*- A test where three participants (two of which are human and one of which is a proposed intelligent machine) would engage in a series of open-ended questions, where the evaluator cannot see the other two participants (Turing, 1950). Their goal was to deduce whether the subject they are questioning is a machine; if they cannot correctly deduce the AI, then the machine according to Turing possesses sufficiently advanced intelligence (Taulli, 2019, p. 13).

Through this paper, Turing began what John McCarthy would further solidify, a notion of an intelligent machine that began the process of detaching itself from computation and mathematics and attaching itself to our hopes, fears, and dreams of the future. It could not be said for certain that things would have turned out differently if the term used did not carry so much implicit weight to us as artificial intelligence does.

Yet, how do we classify these notions today? The most common terminology is that of *Artificial General Intelligence (AGI)* versus *Narrow Artificial Intelligence (NAI)* and that

of *Strong AI* versus *Weak AI.* While many use the terms interchangeably, they represent very different points of discussion in the field of AI. The former denotes a distinction within the field of software engineering, where AGI represents the search for a form of universal learning algorithm which can adapt to any situation or environment and has limitless potential for self-improvement; while NAI denotes a specific algorithm capable of performing a narrowly defined task or type of tasks (Taulli, 2019, p. 4; Zackova, 2015, p. 33). While Strong and Weak AI denote a philosophical classification between machines that are truly able to think, feel and exhibit consciousness and machines that are only able to simulate this type of intelligent behavior; thereby, the terms are not usable interchangeably, although it seems like a sensible expectation that the capabilities of an AGI would be a prerequisite for the creation of a Strong AI (Romportl, Zackova, & Kelemen, 2015, pp. 33-34).

The main issue with having terminology imbued with such weight is that it warps our expectations of what that concept is or can be. As previously stated at the beginning of this chapter, we have been conditioned through centuries of storytelling and representations of what we should expect such *"beings"* to be, and this continues well into the modern-day from the media we consume (such as science-fiction novels, films, etc.) to the promises of the proponents and creators of these tools and machines about how they will set us free from drudge work, how they'll be the end of our mortality or just the end of our species. While it is impossible to separate the social and cultural implications from the term and field, this chapter will continue on to look at AI from the angle of computational intelligence and its current reality of AI as purely algorithmic structures. However, the sociocultural elements of AI will be explored in-depth in another chapter as well. This is important to be able to compare both sides later on in the analysis and discussion phases of this work. In order to better explain the current state of Artificial Intelligence, I must clarify certain terminology such as, for instance, what an algorithm is and what a model is.

What exactly does the term *Algorithm* represent? When speaking in a general system-orientated view, an algorithm is simply a set of instructions utilized in order to obtain a specific end goal. In terms of computation, it is a process of coupling inputs with operational rules to obtain the desired output. While this view is commonly utilized in the field of computer science, there is an argument that even nature runs on *"algorithmic"* principles (Pereira & Lopes, 2020, pp. 25-29). Except, in this case, the source code would be Deoxyribonucleic acid (DNA) rather than zeroes and ones, for example, cells fulfill specific actions (algorithms) based on the rules confined within their DNA instructions when exposed to certain hormones (inputs) (Pereira & Lopes, 2020, p. 46). This way of thinking also leads

us to an interesting limitation of computing as opposed to natural biological *"algorithms."* Unlike biological brains that are capable of exhibiting seemingly contradictory regimes of functioning (such as being under the effects of mind-altering substances or feeling seemingly contradictory emotions at once), computers are only able to exhibit a binary mode of working or not working (Pereira & Lopes, 2020, pp. 25-47). These limitations are inherent in the field since it is both inspired by and opposed to the natural world.

It is from nature that we have derived most of our inspiration for artificial intelligence, thereby also carrying over the inherent limitations and errors found within the very way we function as a species. Such as how the commonly observed effect of Pareidolia (the tendency to perceive patterns where they do not exist, like seeing faces in an electrical outlet, toast, or other objects) carries over quite well into the field of computer vision, where AI can mistakenly flag objects that are not faces as faces, echoing our own natural tendency to do so (Hong, Chalup, & King, 2014, pp. 352-353; Merzmensch, 2020).

In the same vein, we have created several forms of *Machine Learning (ML),* which are very reminiscent of the ways in which we learn (Sanger, 1989, pp. 459-473; Sutton & Barto, 1998, p. 342). Those are the general categories of *Supervised*, *Unsupervised*, and *Reinforcement learning*, each of which carries its benefits, limitations, and issues. None of these approaches can currently fall under any of the previous categories aside from that of NAI since they result in highly specialized tools rather than conscious beings or programs that accurately simulate consciousness. As such, they fall far better under the term of computational intelligence that McCarthy spoke of rather than the more contentious notion of artificial intelligence. This does not discredit their influence or danger; however, it recontextualizes it from a struggle against the ultimate *"other"* into dealing with extremely powerful and difficult-to-understand tools. This distinction is difficult to nail down in a world that often mystifies the reality and capability of such tools, and it also has to contend with deeply rooted expectations and cultural conditioning of our perception of machines (Elish & boyd, 2018, p. 66).

A general understanding of how these forms of machine learning function is required to progress further through this topic to specify the emerging capabilities of these tools. While this is by no means an exhaustive or fully comprehensive list, for the purposes of this work, the three most common forms of machine learning will serve as a baseline for this analysis.

3.1 Forms of Machine Learning

Machine learning is the process by which algorithms gain and improve their ability to discern patterns in a given data set through the establishment and use of statistical models (Taulli, 2019, p. 41). Yet, what are these models? Peter McCullagh defines statistical models as *"a set of probability distributions on the sample space"*(McCullagh, 2002, p. 1225) which alludes to their predictive capabilities. They serve as both the representation and manner of processing data based on certain mathematical principles (for instance, linear regression), and depending on the nature of the data in question they may or may not be used as a representation of a facet of the real world (Huber, 2002, p. 1290). In the case of machine learning, these models are progressively built upon by the machine attempting to codify different relationships within the data it is given so that it can later apply the model to new inputs to give desired outputs.

The most common approaches are the aforementioned *Supervised*, *Unsupervised,* and *Reinforcement Learning*.

Supervised learning allows algorithms to train via metadata labeled data sets (Taulli, 2019, pp. 50-51), for example, if we give an algorithm a set of images of cats and dogs and these images carry metadata that specifies whether it is an image of a cat or a dog allowing the machine to pre-categorize the features it sees on these sets into features that identify cats or dogs in the future.

Unsupervised learning utilizes unlabeled data sets through a process known as clustering, wherein the AI attempts to sort similar examples together (Taulli, 2019, pp. 52-53). As in the previous example, it would analyze the images for similarities (such as the shape of the eyes, length of the extremities, etc.) so that it may group similar images together (which are hopefully distinct and accurate groupings of cats and dogs) to identify any images that we give it in the future and categorize them properly.

Reinforcement learning works on a system of rewards and punishments. The machine is not given any answers upfront; however, you *"reward"* or *"punish"* incorrect results in order to dissuade the algorithm from making the same mistake in the future (Taulli, 2019, pp. 53-54). If the algorithm flags an image of a cat as a dog, we would disincentivize it from making the same mistake in the future by a figurative punishment. Thereby, in theory, the same mistake would not occur again.

Each of these approaches has its own benefits and limitations, as mentioned before. For example, supervised learning can yield very accurate results yet requires an incredible

amount of labeled data, which is not easy to obtain at scale in most cases. Unsupervised learning has a penchant for being very versatile and easy to train but suffers from unreliability and can often learn to see the wrong patterns in the data. Reinforcement learning resides in a middle ground for both difficulty and accuracy, yet it necessitates that the developer has a great number of resources at their disposal from human staff to oversee it to access to high-end computing power.

All three of these forms of learning are reminiscent of the way we learn (Braga-Neto, 2020), especially reinforcement learning, as it is evocative of children being scolded or praised from an early age for their achievements thereby reinforcing good behavior and discouraging bad behavior. Supervised learning occurs in schools where children are given pre-classified information, and Unsupervised learning occurs in our formative first years as we learn to categorize the world around us for the first time (Sanger, 1989, pp. 459-473; Sutton & Barto, 1998, p. 342).

Of course, none of this is strange. It is only natural that we would utilize the ways of learning that are known to us as inspiration for our mechanical offspring. Yet none of these approaches are perfect, not in humans or machines, and can result in hitting many undesirable boundaries and issues. As we see wrong patterns, they can do the same; as we make mistakes categorizing a subject or object, they do it also. These issues were only exacerbated by the development of two very transformative ideas in the domain of artificial intelligence, namely, *Artificial Neural Networks (ANN)* and *Big Data*.

In 1957, Frank Rosenblatt created the *Mark I Perceptron*, a computer program inspired by the functioning of the human brain that worked on the basis of nodes (which he called perceptrons) which would gain or lose their strength and importance (weight) in the network depending on the quality of its outputs (Taulli, 2019, p. 10). This became the first functional *Artificial Neural Network (ANN)*; while it was arbitrary by today's standards with only one layer of data processing, it laid the groundwork for the creation and widespread adoption of *Deep Learning (DL)*. DL is a subset of machine learning which allows an algorithm to tap into patterns that are supposedly nascent in vast sums of data yet imperceptible to humans (Castelvecchi, 2016, p. 22; Taulli, 2019, p. 71).

The creation of Deep Learning has only become possible due to the societal tendency toward digitalization and maximizing convenience, which led to us collecting incredible amounts of data about everything we do, from our shopping habits to health records. The World Economic Forum estimates that by 2025, our global daily output of data will reach 463

exabytes or 497142464512 gigabytes (Desjardins, 2019). This presents us with fertile soil for the development of DL-based algorithms and services.

The pivot towards deep learning revolutionized many industries, creating new opportunities for millions and increasing the efficacy of algorithms in giving solid results. Yet, at the same time, it littered the world with new dangers from misuse to inaccurate or downright dangerous algorithmic conclusions. This was caused by the very structure that deep learning was based on, artificial neural networks much akin to our brains diffuse and encode weights and biases within a wide myriad of different nodes, creating an emergent phenomenon inherent to the complexity of the system itself (Castelvecchi, 2016, pp. 21-22). This makes it very difficult to see how the algorithm came to its decisions, a problem that we have had in researching the human mind for decades (Pereira & Lopes, 2020, p. 6).

However, the problems with utilizing Big Data are not strictly technical in nature; there are also inherent societal issues that feed into this. The proponents of Big Data-driven algorithms claim that they offer unparalleled opportunities and accuracy, yet it comes at a great cost since they can often inadvertently tap into the worst biases and tendencies for discrimination present within our societies (Elish & boyd, 2018, p. 59). One of the reasons for this is that data itself is not neutral or objective, which some of its proponents can discount or gloss over. A data set of prison inmates in the US may hold objective information that a large portion of the prison population is African American. Yet, unless it is treated with due diligence and is taken instead at face value it might lead an algorithmic tool to conclude that African Americans are inherently more dangerous as a group; since it does not take into account the historical reasons for the occurrence of this level of incarceration nor the fact that people are not defined by their racial background (Angwin, 2016). Data and information is never pure or without bias. The story that is constructed with the data can be much more potent than the statistics or data points themselves. Hence it is very important to be wary of the stories constructed in relation to Artificial Intelligence.

And this is where the technical issues come back into the fold. Unless properly managed, such a tool can learn the wrong things; it can learn to correlate dangerous biases such as being dark-skinned with criminality just as easily as it can correlate harmless things like seeing face shapes in electrical outlets.

Yet, how can this happen on the technical side? And why do we have such great issues combating the occurrence of such mistakes? The cause of this is the system complexity itself. A sufficiently complex system, be it a cell, an organ, a human being, or a computer, exhibits properties not found within the sum of its parts. The interplay between these varied

elements (in the case of AI, the structure of the system, and the data it is fed) creates new capabilities within its whole through the process known as *Emergence*. One of the most well-known emergent properties of deep learning machine algorithms is known as *Black Box AI*.

3.2 Emergent Properties of Artificial Intelligence

Emergence as a process occurs within any sufficiently complex system, from neatly arranged atoms of carbon resulting in carbon fiber to the cells in our bodies constituting our organs. An emergent property is any property that arises from the relationships between the elementary components of a complex system that is not present within the components on their own (Aziz-Alaoui & Bertelle, 2009, p. 57; Pereira & Lopes, 2020, p. 42). Thereby, a single atom of carbon does not exhibit the same physical properties as a sheet of carbon fiber, nor do they exhibit those properties if they are not arranged in proper order.

Yet, how does this notion relate to the field of artificial intelligence? Simply put, emergence does not only occur in physical properties but in any relationships between any constituent parts, from people creating states or societies to AI systems creating complex patterns of data processing to arrive at conclusions.

The aforementioned *Black Box AI* is a well-known emergent phenomenon in deep learning-based artificial neural network systems (Castelvecchi, 2016). A black box AI is an artificial intelligence system that, due to its complexity, creates difficulties for individuals or organizations to gain insight into how it arrives at its conclusions. This can occur for many different reasons; For example, the layers of data processing and the abstraction that the inputs undergo through these layers may result in a pattern that would not be discernable to a human expert. Manuel Carabantes states that this is due to the nature of ANN's (Artificial Neural Networks) as *"subsymbolic"* if an AI is symbolic - *"then it is hardly comprehensible by the user, because its heuristic rules, which act as our cognitive biases, are different and also tend to the minimum to explore the whole space of computationally possible solutions."*(Carabantes, 2020, p. 316)*;* yet if it is subsymbolic then *"it is incomprehensible even for its programmer because the operations that transform inputs into outputs are not compatible with human cognition— there are no words, no sentences, no arguments."*(Carabantes, 2020, p. 316). In a sense, due to the abstraction of the data, we lose return information from the system that can be processed by a human way of thinking. There is no discernable feedback on how or why the system gives the answer it gives.

This may also be the result of the very approach to creating these systems, as they are supposedly theory agnostic, we create the architectural systems so that a model may arise through processing large amounts of data, rather than baking in the causational structure behind the issue we're trying to solve (London, 2019). This is in stark contrast to regular statistical models made to process data to gain an expected conclusion from the start. We know how to create a system that may create the solution we want, but not how the process itself will occur when it is in play. A big factor is that we cannot sift through all of the variables in our data that we may perceive as relevant, as the resulting model will reflect regularities in our sample without giving us feedback on the interplay between them. Thereby, any small change to the weights of different nodes may result in a very different model due to the complexity of the system in question, further increasing our difficulties in understanding its underlying processes (London, 2019, pp. 16-17).

According to Carabantes, the black box is a perfectly natural state for this type of complex AI system since the artificial neural networks *"understand"* the world intuitively, akin to our visual cortex, the way of thinking is efficient yet gives little sensible reasoning for its mode of function (Carabantes, 2020, p. 314). He states that if one was to observe the decision-making process of a DL-based neural network in real-time, they would have an aesthetic experience but no inkling of understanding of what is happening from moment to moment as it makes its decisions (Carabantes, 2020, p. 314).

While the emergence of black box AI is a serious problem that requires proper attention and resolution, some researchers believe that the complexity of our systems and the evidence that they exhibit this kind of emergent phenomena is actually a positive development since it signifies the chance for natural artificial intelligence to emerge as well (Romportl, 2015, pp. 214-215; Zackova, 2015, p. 34). Since emergent phenomena are by definition natural and not artificial (even if they originate in an artificial system), if we create an AI system complex enough for an emergent intelligence to form on its own, it would result in a Strong AI, as opposed to creating only a simulation that behaves like a being which would result in a Weak AI. According to Romportl, this Strong AI and its intelligence would ontologically be as natural as ours is since our intelligence is an emergent property of our brains; the term natural within this context relates to whether or not this intelligence manifests itself rather than being a product of design (Romportl, 2015, pp. 214-215).

It may seem like emergent phenomena are chaotic and dangerous. However, by their very definition, they are a product of order. For the human mind, the issue is that the larger picture of how these elements come together and what properties emerge are far beyond our

scope of thinking. Comparatively, it is akin to trying to intrinsically understand what the age of the earth (4.543 billion years) is, while our only personal frame of reference is our own lifetime and experience of the passage of time. Our growth in knowledge or skill in a specific field or task is highly limited by our human lifespans; hence we transfer knowledge through generations. We have never before encountered the sheer learning capacity that an AI can possess, since it can play thousands of years of a game or perform millions of iterations of the same task within a few months, as such it is impossible for us to know what the consequences of this are preemptively (Greenfield, 2018, p. 238).

Yet if these systems are so complicated and difficult to understand for even their creators, why is there such a push for their implementation? The simplest answer would be the quality of results they produce (and thereby their capacity to generate profit); their predictive power is unparalleled by any other form of algorithm since they sacrifice transparency for accuracy (Rai, 2020, p. 138). In a sense, it is a balancing game between the dangers of a technology that is not fully understood and the monetary benefits it can provide. However, due to the dangers of the technology, there are growing efforts to counteract these pitfalls (such as a lack of transparency in decision making, lessening the risk of bias, etc.); a growing body of work has been developed concerning several focal points, namely issues of *Transparency*, *Interpretability,* and *Trust (Carabantes, 2020; Castelvecchi, 2016; Edwards & Veale, 2018; B. Kim, Park, & Suh, 2020; London, 2019; Rai, 2020; Shin, 2021)*. These elements combine into the search for more cohesive frameworks for mitigating the risks of black box AI, such as *Explainable AI* (XAI).

3.3 Explainable AI (XAI)

Interpretability has been a point of concern in the field of computer and system sciences for over 50 years. Yet due to the increasing importance and potential dangers of uninterpretable systems, DARPA, the advanced technological research branch of the U.S. Military launched the *Explainable AI* (*XAI*) project proposal in August 2016, with a goal of producing machine learning techniques that could create explainable algorithms and models (DARPA, 2016; Hansen & Rieger, 2019, p. 41).

While the results of their project are not yet public, the name stuck and became an industry standard. The importance of this move by an influential government body cannot be understated. Opaque systems can result in damage on multiple fronts from loss of human life,

economic downturns, discrimination, or erosion of public trust while at the same time concealing the reasons for the occurrence of these issues.

Explainable AI represents *"the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions. XAI explains the rationale behind the decision-making process…"*(Rai, 2020, pp. 137-138); this may seem to be a rather straightforward and clear definition, yet as usual, within the field of artificial intelligence, nothing is ever quite as simple. To gain a semblance of understanding XAI, we need to understand the concepts of transparency/opacity, interpretability, and explainability further.

Transparency is often described as a desideratum of good, representing a manner of operation open to critical scrutiny; however, it is not desirable in all contexts (Weller, 2019, pp. 23-24). A simple example of this would be the aforementioned classified nature of the research by DARPA; while some may desire it, transparency in this regard may also cause damage by giving powerful technological knowledge to individuals or groups who may misuse it. When related to the notion of transparency in an AI context, it is also manifold in what it may denote from what data we collect or how we utilize it to the underlying system structure itself.

If we contrast transparency with opacity, the interplay between these two concepts becomes clearer. Within the field of artificial intelligence, there are three general forms of opacity, those being *Intentional Concealment*, *Technological Illiteracy,* and *Cognitive Mismatch* (Carabantes, 2020, pp. 311-314).

Intentional concealment is a form of opacity that most commonly occurs in companies or governmental organizations in order to maintain trade secrets and business practices which they do by hiding information about the technology, business, data mining, or any other sensitive information (Carabantes, 2020, pp. 311-312). However, it can also occur as a form of subterfuge to conceal illegal acts or behavior which may cause public outrage. Control for this type of opacity is usually handled on a legislative level. The aforementioned classified research by DARPA is an example of this type of opacity.

Technological illiteracy is a form of opacity most commonly affecting the general population since knowledge in computer and system sciences is an uncommon specialist skill. Similarly to how an individual may know how to operate a smartphone, but not its underlying mechanics or principles, engaging in the ecosystem and understanding it is not the same. Exacerbating this issue is that highly advanced technology such as an ANN is either partially or wholly hidden from the user; they engage with their external access points such as

apps or websites, not the systems themselves, thereby further mystifying the process for the average user (Carabantes, 2020, pp. 312-313).

Cognitive mismatch is the type of opacity that was discussed in the previous chapter that black box AI systems exhibit, where even the experts are unable to understand the inner workings of the system due to its complexity and scale, which at that point begin to lose their connection to humanly understandable concepts or frames of reference (O'Hara, 2020, p. 2).

If these forms of opacity are applied to an example, such as a decision tree, it may be easier to understand how opacity affects a system. A decision tree model is reminiscent of its namesake as a tree that starts from the stem, that based on If-Then statements allows for progression further up one of its branches while limiting our choices based on the previous choice made. For instance, if I have an apple, I could choose to eat it or not eat it. If I eat it, I don't have an apple anymore, only the apple core. Then based on that current state, I could throw the core in the garbage or perhaps compost it. Each step is iterative and limited by the previous choice. This type of system has a very linear and understandable sensibility and is, in general, very transparent.

However, any of the previously discussed forms of opacity could be introduced into this type of system, in turn making it lose its interpretability and transparency. For example, a company that utilizes decision trees could hide the decision tree itself. In this case, it becomes an opaque system to outsiders due to the inability of the public to scrutinize it or observe it. However, this type of opacity is not a property of the system itself (it is not emergent); rather, it is just the lack of access to information (artificially imposed onto the system). Technological illiteracy could be seen as an inherent property to this type of system since there may be individuals who are unable to understand its logic due to, for example, their level of education. If it is sufficiently complex, then the system is highly transparent but not highly interpretable (Hansen & Rieger, 2019, p. 45). While cognitive mismatch could be achieved if the scale of the system becomes too great for a human being to analyze it thoroughly, if it takes a thousand years to check every possible path, then this system is not highly interpretable for you outside of following a few individual decision branches.

It is also important to note that interpretability and explainability are not interchangeable terms. Explainability represents a more general term that also encompasses interpretability, which on its own is not sufficient for understanding black box systems as we require adequate reasoning behind the rationale of the AI for which we need accurate and clear feedback information. When discussing interpretability, we must account for the issue of the human factor (namely that of technological illiteracy); it is important who it is that is

trying to understand these explanations, as a system that spews out highly technical information may not be explainable to the vast majority of individuals and thereby this reduces its explainability (Hansen & Rieger, 2019, pp. 41-42; Shin, 2021, p. 2). Interpretability is a prerequisite for Explainability and cannot be enough on its own in isolation.

If opacity can be introduced to a transparent and interpretable system such as a decision tree, how may we introduce explainability to an opaque system? Unless we have a system that has interpretability and transparency baked into it (which would inherently mean that it is not a black box system), we would have to apply post hoc techniques to gain insight into their inner workings.

XAI is usually split into two general axes of its target and scope. XAI techniques can be *Model-Specific* or *Agnostic* and targeted at the *Global* scope of the model or *Locally* for single instances of prediction; the application of these techniques is meant to align the operations of a black box with a simpler model to turn it into a so-called *"Glass box"* (Rai, 2020, p. 138).

Model-Specific techniques carry constraints meant to increase interpretability within their very structure, while Model-Agnostic techniques use the inputs given to the Black Box along with its results in order to create an explanation (Rai, 2020, pp. 138-140). Both approaches can be applied globally or on a local level.

A *Model-Specific Global Explanations* approach may, for instance, limit the level of abstraction that the data undergoes by reducing the differences between the factors that the algorithm is meant to assess (known as *Monotonicity*) or reduce the number of inputs that go into the model (known as *Sparsity*) (Rai, 2020, pp. 138-139).

Whereas a *Model-Specific Local Explanations* approach would focus on explaining the results of a specific example through tools like heatmaps which would highlight which elements of an image contributed most to its predictions (Rai, 2020, pp. 138-139).

Model-Agnostic Global Explanations attempt to approximate what a simpler, more interpretable model as close as possible to the Black Box would be like, for instance, through decision trees based on IF-Then statements (Rai, 2020, pp. 139-140).

Whereas *Model-Agnostic Local Explanations* are similar to their Model-Specific counterpart but focused on common denominator interpretable components inherent in all ML systems (Rai, 2020, pp. 139-140).

These approaches, just like different forms of machine learning, are varied in their uses and limitations. Model-Specific techniques are more expensive yet more accurate, while

Model-Agnostic techniques are more difficult to apply and generally less accurate yet act as generally applicable tools.

Yet, what is the actual value of XAI? What is its purpose besides illuminating a technical issue? The main reason for the existence of frameworks such as XAI is to facilitate the establishment of trust in AI-driven systems not only with experts but also with the public (Jacovi, Marasović, Miller, & Goldberg, 2020, p. 1). This is important not only because of the actual complexity of the technology itself but also from the social expectations of AI and the often disingenuous media frenzy surrounding the technology itself and how it is presented to the public; rather than the idea of computational intelligence, it is sold as the cure for all issues in our lives and is heavily anthropomorphized (Troshani, Rao Hill, Sherman, & Arthur, 2020, p. 3). XAI as a concept is mostly concerned with issues such as the opacity caused by cognitive mismatch, and some critics believe that this is a problematic assumption to make for the general population, that the lack of trust is an issue of the public not comprehending the technology (which at the same time is mystified and its capabilities are blown out of proportion in the media space) rather than its pervasiveness or inadequate legal safeguards against its abuse or incorrect application (Knowles & Richards, 2021, pp. 1-2).

Trust in AI systems can be warranted or unwarranted, intrinsic or extrinsic, and is often based on the expectations of the individual that they are stepping into a sort of contractual trust with an organizational entity; however, Human-AI trust is different from interpersonal trust, yet since it is anthropomorphized unlike most technologies (due to the perception on the part of the user that it is a technology endowed with intent or agency) it is not treated as just a tool, rather it exists at a sort of impasse between the categories of a tool and an entity (Jacovi et al., 2020, pp. 1-2).

Whether or not trust in an AI is warranted is closely related to its trustworthiness. However, one must not mistake being trustworthy and being trusted as the same -*"Trust can exist in a model that is not trustworthy, and a trustworthy model does not necessarily gain trust"*(Jacovi et al., 2020, p. 4). Here we see the distinction between warranted and unwarranted trust; it is warranted if the subject is considered trustworthy and unwarranted if it is not. Yet this notion is closely related to how we communicate what an AI is, how it works, what it is not, etc. The anthropomorphic language that is often used to bring a difficult-to-understand technical concept to the public does not communicate the trustworthiness of these technologies to the wider population; instead, it may be damaging it, alongside scandals such as Cambridge Analytica. Some researchers believe that AI suffers from a great *"trust deficit"* from the outset and that it will be very difficult to create public

trust in AI regardless of initiatives such as XAI, which are more useful to the experts than the public itself (Knowles & Richards, 2021, pp. 4-7).

The aforementioned intrinsic and extrinsic forms of trust play into this dichotomy:

Intrinsic trust can only be established via the understanding of the underlying principles of the technology or its decision-making process, and it is heavily dependent on excising both technological illiteracy and cognitive mismatch from the system (Jacovi et al., 2020, p. 5). Knowles & Richards argue that the general public does not need intrinsic trust in AI, rather in their view, having robust systems of regulation and transparent experts who would be a stern source of extrinsic trust are of greater importance (Knowles & Richards, 2021, p. 9).

Extrinsic trust could be seen as a general inverse of intrinsic trust, as it originated from external sources such as proxy agents (experts within the field, governmental organizations, trust in the credibility of the purveyor of the technology) or performance/test indicators from the results of the AI itself being satisfactory (Jacovi et al., 2020, pp. 6-9). Of course, neither form of trust is easily established when it comes to novel occurrences in people's lives; however, XAI does generally seem to be a step in the right direction since it can allow certain individuals to both build intrinsic trust themselves and through the discourse on creating more transparent and responsible systems can also bolster extrinsic trust by proxy (Shin, 2021, p. 3).

Knowles & Richards also believe that we must build what they call trust in *"AI-as-an-institution"* since it would be disingenuous to frame the discussion around trust in AI as individuals earnestly interacting with single clear instances of AI; rather, the public experiences AI as a pervasive evermore intrusive concept and as such requires the establishment of general trust in the idea of AI in being a part of our lives (Knowles & Richards, 2021, pp. 1-4).

If we are beginning with a *"trust deficit"* in regard to Artificial intelligence, we must look into the reasons for this position. From cultural conditioning to the imposed ontological standing of other beings in relation to humans, there are many possible culprits for the source of this deficit. Another important question is what happens when this trust is broken, especially in regard to our perceived experience of the machines as machines rather than as intelligent thinking beings. This is the moment that can fill certain individuals with a unique form of unsettling sensation known as the Uncanny Valley. What happens when our long-established expectations (and promises) meet our complex artificial systems whose very logic is incomprehensible to us?

4 The Uncanny Valley

In 1970, Japanese roboticist Masahiro Mori proposed the existence of one of the most contentious yet influential phenomena in the field of robotics. He proposed that the likeability of a robot was closely linked to its resemblance to humans with a near-linear progression in affinity upwards to a certain point at which the aesthetics of the robot would delve into the territory of realism; as this point is approached, the affinity would turn into eeriness unless the robot became nigh indiscernible from a healthy human (Mori, MacDorman, & Kageki, 2012, pp. 98-100) *(see fig. 1 and 2)*. The term Mori used for this phenomenon was "*bukimi no tani genshō"*(Mori et al., 2012, p. 98), which was translated into an English publication as *"The Uncanny Valley Phenomenon."*

Figure 1. Sophia (ITU, 2018)

Figure 2. Kaspar (Hertfordshire, 2005)

This idea would cement itself as a point of interest in many scientific fields from robotics to human-computer interaction (HCI), as well as into popular culture and mainstream lingua franca as a catch-all term for the creepiness of badly-designed robots.

Mori's hypothesis is that the phenomenon was caused mainly by the dissonance between the familiarity of (or affinity toward) an object and those expectations being shattered upon the realization of the actual nature of the object (Mori et al., 2012, pp. 98-100). He gives the example of a prosthetic hand which in the dark may seem like a normal human hand, yet touching it would result in confusion and discomfort due to its cold rubbery touch instead of the expected familiar sensation of human skin. A contributing factor in his view was the addition of movement, which may amplify the curvature of the uncanny valley, thus deepening the discomfort an individual would feel (Mori et al., 2012, pp. 98-100).

Over the years, many have criticized the idea of the uncanny valley, while others have expanded upon it and looked for better justifications for this sensation. This has resulted in the formation of other subsequent theories and frameworks.

The most refined theories borrow from established concepts in the field of psychology, namely the *Categorization Ambiguity* and *Perceptual Mismatch* hypotheses.

Categorization ambiguity as a source of the uncanny valley effect centers on the issue of mentally categorizing realistic artificial beings as real or artificial. It focuses on the notion that there is a *"categorical boundary"* between two categories (such as real human and robot) where objects or subjects that straddle the line or are otherwise close to it are difficult to categorize concretely into either of the two options (Kätsyri, Förger, Mäkäräinen, & Takala, 2015, pp. 5-6). This, of course, implies that there are such clear-cut categorical boundaries in general and reflects on the commonplace tendency in society to categorize people within distinctly defined groupings. This type of thinking also leads to the otherization of individuals or groups that do not fit neatly into these categories, and thereby the idea of categorization ambiguity implies that any entity that may not be neatly sorted is inherently creepy.

In regards to the uncanny valley, this prolonged uncertainty while attempting to categorize the subject is the supposed source of the discomfort or eeriness (Kätsyri et al., 2015, pp. 5-6).

The perceptual mismatch theory, on the other hand, is less rigid and conceptualized on a form of continuum wherein the cause of the uncanny valley is the observation of subtle inconsistencies in the human-likeness of the subject, which violate our expectations about the true nature of the *"entity"* in question (Kätsyri et al., 2015, pp. 6-7). While this does seem like a more cohesive idea in regard to artificial entities, it still carries certain societal implications, such as the idea that with prolonged scrutiny, subjects may reveal their true nature as the other. This is of particular interest when considering real-life occurrences of trans or gay panic defenses where perpetrators of violence claim that they were being deceived about the gender or sexuality of their victims. The idea of perceptual mismatch is sometimes split into two: the *Inconsistent Human-likeness* and *Atypicality* hypotheses. The main difference is that the atypicality hypothesis is focused only on "human-like characters" and distortions found in our expectations of what an average human is (Kätsyri et al., 2015, pp. 6-7). However, this also brings forth discussions of what this *"average"* human that the entity is compared to is.

These are the main theories that this chapter will utilize in order to explain the uncanny valley concept further. Yet, it would be remiss not to briefly mention the so-called

"simple" uncanny valley theories. Those being: The *Naïve hypothesis* that states that any kind of manipulation can induce the sensation of the uncanny, which is seen as overly simplistic and neglecting of the fact that not all forms of alterations are relevant in causing the phenomenon (Kätsyri et al., 2015, p. 4). Thereafter, we have the *Morbidity hypothesis,* which originates in the original uncanny valley theory and relates to the notion that these feelings stem from deeply rooted morbid associations to corpses or zombies being projected onto robots (Mori et al., 2012). Lastly, there is the *Movement hypothesis* which focuses on imperfect human-like movement eliciting and amplifying the UV curve further (Kätsyri et al., 2015; Mori et al., 2012).

Yet, these theories are often seen as lacking in nuance or substantial empirical evidence. This thereby infers that their origin, that is, the uncanny valley conception as defined by Mori is also highly simplistic or even unscientific. However, with the advent of these refined theories, which are more susceptible to both empirical study and general academic scrutiny, the topic has been vigorously debated for over fifty years. This lends more credence to its importance and influence.

Mori never meant to create such a fervent debate over the topic; in his mind, this was created from a simple observation of a sensation he perceived during his work, and as such, he wanted to share it with his community of robot designer peers (Kageki, 2012, p. 106).

Of course, this cannot be taken as a capitulation for not scrutinizing the idea itself, yet I must state that my very wish to explore this subject was born directly from my own experiences of the uncanny in relation to artificial beings, be they videos of robots with unsettling expressions or virtual characters in games that due to errors exhibit disturbing changes in their body or position and thereby break my suspension of disbelief. In a sense, this chapter also serves as a way to explore and recontextualize my own feelings on the subject. This is why it is very important to look into the criticism of the idea as well as the ways in which others expanded upon it in order to reach a greater understanding of this subject matter.

4.1 Criticism and Empirical Studies

It would be difficult to summarize every area of criticism that different research communities delved into with the uncanny valley as posited by Mori. One highly relevant factor in the debate being whether we are focusing on the original or derived hypotheses (Kätsyri et al., 2015, p. 2).

When looking at the original concept, the issues are manifold, from its highly speculative nature to its vague definitions of key metrics and gauges for the area where the valley is; yet the most interesting thing about the criticism of the uncanny valley is how it became fertile soil for people who through critique expand upon rather than dimmish the idea itself.

One of the most salient critiques (and thereby most interesting additions) to the uncanny valley hypothesis is the focus Mori placed on human-like robots, neglecting the notion that virtual characters or artificial intelligence could exhibit the same influence on the observer (Draude, Aylett, & Michaelson, 2011, p. 321). This is exacerbated by the focus Mori placed on the perception of the touch or materiality in relation to the uncanny valley (Mori et al., 2012).

Yet many forms of digital media have had the label of uncanny hoisted upon them, such as the film *Mars Needs Moms (2011)* for their usage of nigh unsettling 3D models of humans, or the game *Mass Effect: Andromeda (2017),* which due to its myriad of bugs on release resulted in stunted facial animations and eerie movements of characters which suggests that many forms of human-made objects or subjects may slip into the uncanny.

Another criticism of the hypothesis is not taking into account prolonged interaction with the subject in question, which may cause the eerie sensation to dissipate over time as individuals become accustomed to it (Rhee, 2013, p. 306). Mori saw the uncanny valley as something to be avoided via the purposeful change in robot design ethos by focusing on non-human design (Mori et al., 2012). In contrast, other researchers propose that we should not capitulate to the sensation by avoiding it but that we must instead create more human-like artificial entities in order to deconstruct the very issue through constantly encountering it (Romportl et al., 2015, pp. 134-135). This idea of prolonged interaction echoes back to the notion of otherization related to the refined theories of the Uncanny Valley, as well as research in social psychology that tackles implicit bias and discrimination through increased interaction between in and out-groups as one of the most effective methods to reduce discrimination and hate (Devine, Forscher, Austin, & Cox, 2012, p. 8).

This prolonged interaction could, in turn, *"inoculate"* us against this sensation in a future world saturated with robots and AI. Notwithstanding exceptional cases, of course. However, if we interacted with such entities on a daily basis, our familiarity with them could cancel out the initial hesitancy. In some cases, familiarity is a two-way street, such as with the robot Kismet who behaves differently based on whether he knows a person (like its creator) or is first interacting with them (Rhee, 2013, p. 306).

This habituation also factors into many other angles used to explore this topic. Is the capacity to experience the Uncanny Valley something that we are born with? Or is it learned? Mori believed that this was an evolutionary survival instinct (Mori et al., 2012). Whereas Brink, Gray & Wellman, in their attempt to answer this question found evidence that very young children (under 4) do not experience robots as uncanny and seem to attribute to them vastly greater capacity sentience and capabilities while older children perceived them as uncanny if their behavior did not match their expectations (Brink, Gray, & Wellman, 2019, pp. 1203-1211). This indicates that the sensation is more likely to be developmental rather than evolutionary in nature.

This notion is further bolstered by some researchers finding evidence that older subjects (average age of 60-65) experienced either a diminished level or no eeriness whatsoever in relation to human-like robots; some of the participants even preferred them (Tu, Chien, & Yeh, 2020, pp. 389-390). Whereas young adults consistently showed a preference toward non-human robots (Tu et al., 2020, pp. 390-391).

This could imply perhaps that there is also an element of cultural conditioning since younger generations continuously encounter depictions of robots and AI in their entertainment, such as R2D2 in *The Star Wars* franchise or the T100 in the *Terminator* franchise, which could shape their expectations and perceptions of non-human and human-like robots in general.

Some attention has also been given to the personalities of the test subjects themselves. Mainly in relation to their avoidance of novel experiences or stimuli, which seems to indicate how likely or how strongly individuals may experience the sensation of the uncanny, with a positive correlation between the novelty avoidance trait and the intensity of the valley (Sasaki, Ihaya, & Yamada, 2017, pp. 2-10).

An important caveat that must be specified is that none of the previously stated research utilizes a pure form of the original theory; most of them either follow the categorization ambiguity or perceptual mismatch hypotheses. This, in a sense, both discredits and validates the idea, the sensation seems to have empirical backing, but the original theory was not fully formed. These refined theories shift the uncanny valley from being an intrinsic property of the objects in question to a complex interplay between us, our expectations, cultural conditioning, and the machines.

However, even the empirical evidence exploring these two conceptions of the uncanny valley isn't conclusive. Metanalyses of multiple studies found evidence in regards to perceptual mismatch, while categorization ambiguity remained inconclusive (Kätsyri et al.,

2015, p. 12); yet other studies found evidence for the categorization ambiguity hypothesis while stating that it could also be attributed to perceptional mismatch as well (Sasaki et al., 2017; Strait et al., 2017).

Among the reasons for these confusing results might be that the very nature of the uncanny valley is not predilected on a singular source, but as previously stated, it might be a complex web of different issues and stimuli that induce it. This can also be seen in the wide range of fields that the theory is explored in, from human-computer interaction to gerontology. And each of these fields utilizes different methods to arrive at their conclusions.

There is yet another factor that I have yet to account for. What is the uncanny? Understanding the conceptual basis for this feeling might help us get closer to a concrete image of the uncanny valley.

4.2 What is the Uncanny?

Perhaps the most seminal text in regard to the uncanny is Sigmund Freud's 1919 work of the same name. To Freud, the uncanny (*"Unheimlich"*) represented the opposite of what is familiar, known, or belonging to the domain of the home (*"Heimlich"*), thereby the Uncanny denotes that which is unfamiliar (Freud, 2004, p. 418). Of course, not everything unfamiliar or new is disturbing; to Freud, a certain *"something"* must be added in order to make something Uncanny (Freud, 2004).

It is interesting to consider this as it relates to the uncanny valley, as Mori's conception was firmly entrenched in the field of aesthetics. This echoes Freud's notion that while the uncanny isn't just aesthetic in nature, it is very dominant in that field (Windsor, 2019, p. 53).

Yet what may be this *"something"* that must be added to make something uncanny? According to Friedrich Schelling, things become uncanny when something that should've remained hidden or secret surfaces; the uncanny thus represents a revelation of the true nature of something; this further echoes Erns Jentsch's idea that uncanniness relates to intellectual uncertainty, especially in regards to whether an object is or is not, in fact, animate (Freud, 2004, pp. 418-421). These notions also imply a deterministic view that there exists such a thing as *"true nature"* that could be unearthed; when looked through a more relativist lens, this idea wavers as what something is or is not depends on circumstances, such as who is observing it, what events are occurring or how the object/subject is scrutinized.

When this definition is compared to the categorization ambiguity theory, it is very easy to notice the parallels. This intellectual uncertainty that Jentsch spoke of in a sense denotes the anxiety related to being unable to categorize an object into its *"proper"* group. While perceptional mismatch lends itself well to the idea of revealing, as we perceive the imperfections or inconsistencies in the subject of our view, they slowly reveal their true nature to us as not human; or so it may seem.

It is through this revelation that the sensation of the uncanny instills itself into us. Yet, it is unclear whether the sensation itself is a mood or only an emotion, as emotions would be directed at specific subjects while moods are all-encompassing (Windsor, 2019, pp. 55-56).

This distinction may be important to understand the source and form of the uncanny that we are dealing with in different circumstances. While some people likely have a general aversion or dread connected to the idea of robots, the uncanny valley as an experience seems to be more directed towards particular objects of our focus. Less so as a form of fear in the sense of concrete danger to us, and more so as anxiety in relation to unknown threats or questioning the perceived true nature of an object as previously stated (Windsor, 2019, p. 57).

There is a great deal of uncertainty concerning artificial intelligence and robots, from fears of job losses to more doom-laden predictions of AI being the potential downfall of humanity. When this is coupled with decades of science-fiction entertainment that conditions us with different expectations of what AI is, will be, or should be, it is not difficult to surmise that these depictions will seep into our expectations for their real-world counterparts.

These expectations that we form might also stem from a deeply rooted perception of machines as the ultimate *"other,"* a group so distant from human experience and nature that it cannot be treated or perceived as anywhere close to us as a species. After all, their existence is predicated on the idea of them being purpose-built creations meant to fulfill tasks; their function is their existence. This notion of servitude as inhuman tools is reminiscent of the historical justifications used in an attempt to justify the enslavement of people throughout history, from the idea of other groups as *"sub-human"* to class-related struggles and indentured servitude based on social status. However, many may see the idea of equating machines in any capacity to humans (even if they were highly intelligent) to extend them rights as nonsensical. However, one must not forget that similar arguments were and still are applied to people all around the world to justify horrible infringements of their rights as human beings.

Yet as we create more and more technologically advanced *"entities,"* we risk getting closer to that domain of ethical issues that science-fiction has dealt with for decades. Could

our robots be sentient, filled with emotions or hopes and dreams? This is an entertaining idea to explore in fiction, but in reality, when we reach that point, we will have to contend with great problems that need solving.

This idea challenges the *"sanctity"* of being human as well as our place in the world. For many individuals, this may cause an existential anxiety in regards to the notion that what it is to be human is being demoted or destroyed by ever-encroaching technological advancement (M.-S. Kim, 2019, pp. 322-327). This is akin to how Heliocentrism and Darwin's theory of evolution shook our egotistical view of human exceptionalism to what some perceive as lower status (M.-S. Kim, 2019). These changes in our ontological standing in comparison to the rest of existence have changed many things, including, for instance, our perception of animal rights, and may, in turn, change our perception of machine rights when the time comes.

This danger for our status as humans might be the underlying cause of this sensation of the uncanny. As a machine that was meant to reflect us and display characteristics close to us reveals its *"true"* inhuman mechanical nature, we recoil at this realization. It is not close to us; it is a machine masquerading humanity and, as such, is a violation. When observed from a categorization ambiguity lens, this process leaves us trapped between having to decide on where this entity belongs, amongst us or the machines. Yet if we look at it through the view of perceptual mismatch, it is a drawn-out dread-inducing process of realization about what it actually is that we are facing.

But what happens when the source is not the physical representation itself? What if the very idea that a robot might feel or be in control might induce the sensation? As previously stated in the text, Jentsch believed that this feeling originated in the dissonance of being unable to tell whether an object is animate or not (Freud, 2004, pp. 418-421). This does not sound very much like an aesthetic experience, more so as perceiving a mind or being.

4.3 Uncanny Valley of the Mind

The *Uncanny Valley of the Mind hypothesis* originates in the criticism of the original uncanny valley concept; it was first posited by Kurt Gray and Daniel M. Wegner in 2012.

In their view, the uncanny valley is an experience that originates not in the aesthetics of a machine but in our ability to perceive and attribute a mind to it (Gray & Wegner, 2012). They split this attribution of the mind into two parts, the experience of perceived agency (the

ability to act independently, make choices, and execute plans) and perceived experience (the ability to feel emotions and be aware of oneself) (Gray & Wegner, 2012).

While this chapter has previously reviewed an obviously present and strong connection to aesthetics, that cannot be the only realm which this sensation can occupy. Even more so, the previously stated conceptions of the uncanny itself refute this notion, as it seems that the aesthetics of an object only serve as a conduit for a more fundamental reason for its existence. The reason is the dissonance between our expectations and reality, such as not being able to define whether an object is alive or not, which is directly connected to this notion of attributing a mind to an object, be it through perceiving experience or agency.

Gray and Wegner believe that the capacity to experience is more fundamental to the human condition, that our high capability for both agency and experience is what separates us from other animals. Thereby in their view, an AI with a high perceived capacity to experience would be more uncanny to us than one with a high degree of agency (Gray & Wegner, 2012, pp. 125-127). Encounters with such entities would violate our internalized expectations of what a machine is or should be capable of.

Of course, one can also see a modicum of human exceptionalism that underpins their view of agency and experience, as if animals were incapable of feeling complex emotions. Nevertheless, in their tests and in the subsequent work of other researchers, there have been positive correlations towards validating this idea (Appel, Izydorczyk, Weber, Mara, & Lichetzke, 2020; Stein & Ohler, 2017; Van der Woerdt & Haselager, 2019). Yet, there is a large caveat to their empirical work, as it is mainly based on vignettes and descriptions of robots, as today we do not have sentient machines to utilize in such tests. However, as the focus is on the perceived mind of a machine, rather than actually encountering one, it doesn't immediately discredit their work.

While Gray and Wegner found that robots with perceived agency aren't eerie, other researchers have found that while they are not as intense, these types of machines or AI are also able to elicit an experience of the uncanny (Appel et al., 2020, pp. 275-278).

How well do these two attributions of mind lend themselves to the aforementioned categorization ambiguity and perceptual mismatch hypotheses?

In the case of the former, the issue seems to stem from defining whether a being is independent or emotional in rather rigid binary terms, which isn't a very clear-cut barrier, nor can it be easily done. In comparison, the continuum-based perceptual mismatch fits much more neatly into this framework as a process of unveiling or revealing the mind of the machine to us. Inducing a slowly creeping realization that the mechanical being in front of us

might be able to think and feel, which may instill existential dread related to our expectation that it was just a very interactive programmed object.

If this notion of revealing is related back to the notion of machine *"otherness"* discussed earlier in this chapter, the uncanny valley of the mind would seemingly serve as a form of threat avoidance, a way to protect either the ego or personhood of an individual against the danger of being superseded by machines rather than as a fear of a physical threat. For if an AI exhibited a highly advanced or sophisticated mind beyond our expectations, it could induce a severe form of cognitive dissonance on our present worldview of both what it means to be human and machine (Stein & Ohler, 2017, p. 45).

At the moment, there aren't any machines that could cause this by displaying actual sentience, yet the mere perception of a twinge of this realization may induce the feeling that something is just not right.

This may be linked to the idea that the most important element in our interaction with robots isn't their actual capabilities but rather the way we experience them (Gahrn-Andersen, 2020, pp. 1-8). As to an average observer, high technology such as AI is inherently black-boxed, giving them little input or awareness of how they function. This is much akin to how one knows how to use a smartphone but not its underlying processes or technology. Thereby it is irrelevant whether the machine only seems autonomous/sentient or whether it really is; this also links to the discussion on technological illiteracy.

Indeed, it seems that people have an inherent tendency to anthropomorphize artificial entities and infer upon them a sense of responsibility and personhood if their actions cause damage or harm (Van der Woerdt & Haselager, 2019). This goes so far as to ascribe individuality and agency to an AI or robot if it displays a lack of effort, which results in its observers deriding its actions or behavior (Van der Woerdt & Haselager, 2019).

If people aren't aware of the manner in which an AI functions and are conditioned by decades of entertainment and fiction on the subject, their expectations of what an AI or robot is supposed to be are twisted from the outset. Provided that there is a general tendency to anthropomorphize machines and infer upon them being, this seems to be the perfect breeding ground for the uncanny valley.

This issue is exacerbated by machines not processing information as we do. Whereas human rationale is based on interpretation and is context-sensitive, machines only process representations of things, and their logic is rigidly rule orientated (Draude et al., 2011). This severely limits our communication with machines turning it into an arrangement of signal processing. And in that process of abstraction, we lose a large part of what makes the

information comprehensible to humans. In a sense, the information we gain from them has to be translated back into something that we can interpret. This is why robots and digital avatars are prime tools for bridging the gap, as they transform this information into behavior and actions more understandable to humans and may even assist in deconstructing the black box AI (Draude et al., 2011, pp. 321-322).

But small errors or discrepancies in this process may result in the imperfections argument detailed within the perceptual mismatch theory, resulting in stunted or irregular behavior, which may induce the uncanny valley effect.

Yet if this is the case, wouldn't all stunted AI behavior result in this sensation? Interestingly enough, some people prefer AI that exhibits simulated emotions or mood states like Siri or Alexa (Stein & Ohler, 2017, p. 44). What makes these displays of AI emotion endearing whereas other tests result in eeriness? The simplest argument for this would be the awareness of the status of the AI; these emotions aren't proof of sentience. They are programmed and highly limited to generalized comical banter. There are no qualms around whether Alexa is scheming to destroy the human race; thereby, the very element supposedly required for the uncanny valley – that is, the uncertainty around the nature of the object in question – is not present.

Through the many arguments for and against the different conceptions of the uncanny valley, several key elements arise as the potential building blocks of the phenomenon. Those are uncertainty about the nature of the object, the violation of our preconceived notions and expectations about the objects of our scrutiny, and the threat to our ontological standing as human beings. These elements are present as underlying forces in all conceptions of the uncanny valley, be they in regard to aesthetics of artificial beings or attributing and perceiving a mind within them.

However, what may happen if the attribution of mind is exacerbated by some very inhuman yet intelligent behavior? Since it is established that machines do not think in the same way humans do, would a highly advanced AI induce the uncanny valley phenomenon if its actions have uninterpretable origins or logic to us? If it behaved like no human would ever behave or made choices that seem to have no actual rationale behind them to any human yet seem very intelligent and calculated. Would it inflame our dread about the future due to the often-prophesized AI-induced human obsolescence? A conception of this extreme version of the uncanny valley of the mind is the subject that I will discuss in the next chapter.

5 Uncanny Logic

The previous two chapters detail two different sides of the debate related to the experience of Artificial Intelligence, the former is focused on the technology itself and its many varied elements and the latter on the perception, unease, and cultural imaginaries related to the notion of intelligent or *"living"* machines through the lens of the Uncanny Valley. Their purpose is to assist in the familiarization with the subject matter of this text as well as lay the theoretical groundwork for the defining of the concept of *Uncanny Logic*. The inspiration for this comes from the statement from Adam Greenfield's Book on the way certain observers of AlphaGo's matches against Lee Sedol felt watching it play -*"But there was something almost numinous about AlphaGo's play, an uncanny quality that caused at least one expert observer of its games against Lee to feel "physically unwell."."(Greenfield, 2018, p. 238),* this odd description created a wish for defining this idea of intelligent AI behavior so abnormal to human eyes that it could induce great unease. While the concept of the Uncanny Valley in relation to the aesthetics of robots has been greatly explored, and the Uncanny Valley of the Mind hypothesis has gained ground in regard to the perception of mind in machines, I believe that there is a lack in describing this specific notion of uncanny logic as a sub-set of Uncanny Valley of the Mind.

To arrive at a more cohesive idea of uncanny logic, this text will require starting from a more simplistic statement and building upon it; as such, the working definition, for now, will be – *Uncanny Logic is a subset of perceived agency induced uncanny valley of the mind originating from complex opaque artificial intelligence decision making and behavior.* To better formulate this definition, as well as elaborate the reasoning for this stance, I will draw parallels between the two theoretical sections searching for the elements that may constitute this concept.

To define the concept further, we will cut the working definition into the two general parts that constitute it; however, they will seep into each other during the analysis:

1) Perceived agency induced uncanny valley of the mind.
2) Opacity and complexity in AI decision making and behavior.

The previous chapter details the main elements that exist in all conceptions of the uncanny valley, those being the uncertainty about the nature of the object in question, violation of preconceived notions or expectations about the object of scrutiny, and the threat to the ontological standing of humans. If these elements are taken as a requirement for

conceptualizing the uncanny valley, then the concept of uncanny logic would exhibit all of them in relation to AI systems while also maintaining the focus on the perception of mind through agency as the lens through which they are expressed.

The chapter on AI theory details the other half of the requirements. Namely, the concepts of emergent phenomena through complexity and the two most technologically driven forms of AI opacity; those being cognitive mismatch and technological illiteracy.

Mediating influences between these two groupings are carried out by issues of language that anthropomorphizes the tools themselves, cultural conditioning, and trust in machines.

With that generalized shorthand dissection of the working definition, the process of defining uncanny logic is set in motion.

5.1 A Nebulous Idea

Long-standing cultural imaginaries of intelligent machines coupled with the term Artificial Intelligence are partially at fault for the state of the field of AI research and development today (Heffernan, 2020, p. 93); immersing the public in the expectation of a world inhabited by not only *Artificial General Intelligence* (AGI) but also thinking intelligent beings tied to the notion of *Strong* AI (SAI) with experiences, agendas and a drive for self-preservation (Romportl et al., 2015, pp. 33-34). While these machines do not exist today, the notion that they might someday be possible leads to ideas of technological utopias/dystopias and even influences our current experiences of our (by comparison) primitive technologies. Thereby, even the slightest perceived inkling of this future could induce great unease in individuals via the perception of a mind within the machine, be it one that can act freely of its own volition or experience the world as we do (Gray & Wegner, 2012, pp. 125-126). As previously stated, we may never know if we could have been spared some of these issues if John McCarthy coined the term *Computational Intelligence* instead to denote his view of AI (Heffernan, 2020, p. 93); while doubtful that it would have changed much as our relationship to our technology and fears of it running amok have deeply rooted historical roots, it may have eased the stress on this particular idea of machine intelligence as a function of algorithmic computation. Rather than intelligent beings, they would simply be computers making calculations to solve mathematical problems; while by all means still an impressive feat, we are not in terror of the computers on our desks, for they are just tools.

However, thanks to certain market forces and interest groups (namely purveyors of said technologies), the image presented to the public is warped, it is intrinsically tied to the view that this technology has the potential to be the greatest feat of our species, which by all accounts it could, but not now, nor soon as our mistaken predictions show (Armstrong & Sotala, p. 28). The language that is used often espouses glorious potential and ascribes forms of individuality or capabilities far beyond that of a human, all of which is neatly packaged into an anthropomorphized view of AI as a distinct grouping of programs transcending their status as mere tools and delving into the domain of beings (Elish & boyd, 2018, p. 66). Yet when this is coupled with the tendency people have to anthropomorphize artificial entities to the point where they can ascribe responsibility and the status of personhood (Van der Woerdt & Haselager, 2019), this quickly becomes a spiraling feedback loop of hype between both the public and the creators of AI that sets the bar of expectations (and fear) incredibly high.

This is the reason why a notion of *"AI-as-an-institution"* by Knowles & Richards could be of use, as the average user simply is not aware of the scope, form, or influence of AI in the most realistic technical sense; instead, it is experienced as an idea that has a sort of notoriety as an ever-present and intrusive facet of modern life (Knowles & Richards, 2021, pp. 1-4). This element of the experience of the technology itself is very dangerous since, in some regards, the actual reality behind the capabilities of the technology is irrelevant to a user if they experience them differently (Gahrn-Andersen, 2020, pp. 1-8) in this regard utilizing the more realistic notion of computational intelligence when discussing AI will be of little use in practice when dealing with the public, the public does not experience computational intelligence, they experience artificial intelligence, with all the weight the term carries.

Of course, this does not mean that experts shouldn't pay attention to the issues, but also to include the subjective experience of the technologies in their process, regulations, development plans, etc.

Yet why is the experience of AI so drenched in unease? If we look back at the discussion related to the notion of the Uncanny, it is not a concrete fear resultant from physical danger; it is a form of dread or anxiety at the prospects of the truth of what it is that we are experiencing or observing (Windsor, 2019, p. 57). If we anthropomorphize the current state of AI and if it is perceived as a sort of permeating influence, it could inherently lead some individuals to feel like the technology is already out of control. When this is coupled with our present cultural expectations and imaginaries of living machines through history, which is steeped in fears surrounding the generation of artificial sentience (Elish & boyd, 2018, p. 62), we are placed into a position of reflecting upon the only experience we have had

with such concepts and that only exists in science-fiction. Many of those stories serve as cautionary tales, often related to fears of loss of humanity, extinction, or even the demotion of our very standing in the world; all of which can feel like an "attack" on the sanctity of being human and thereby induce existential dread over ideas such as robots being on equal standing as us or having rights (M.-S. Kim, 2019, pp. 322-327).

While Knowles & Richards claim that the deficit of trust in AI comes from a lack of understanding of the technology by the layman coupled with scandals (Knowles & Richards, 2021, pp. 4-7) from Cambridge Analytica to deaths by autonomous vehicles, an equally strong and credible stance would be that these issues began far in our collective past, and are mediated by those cultural imaginaries, fears, and hopes (Elish & boyd, 2018, p. 62).

An important part of this is also the current state of the technology; while it may not compare to the science-fiction technology that we are used to, it is well beyond anything that an average person several centuries ago could have truly expected to be a reality. This is where the importance of emergence and complexity comes into play. Today's *Artificial Neural Networks* (ANNs) possess enormous capacities for data processing; when this factor is coupled with the mystifying and anthropomorphizing language often used to describe AI's it is not difficult to see how easily the description of an AI inspired by the structure of the brain (Taulli, 2019, p. 71) may connect to the aforementioned expectation (or fear) of nascent Strong AI residing somewhere in our systems ready to take over.

Yet if we have a sense of these expectations, where does uncanny logic fit into the discussion? If combined, the aforementioned discussion on the experience of AI and cultural connotations leads to an element I believe is very important for the defining of Uncanny Logic; and that is the concept's predication on a feeling of unease in encountering a perceived intelligent inhuman entity that parries or surpasses our level of intelligence or capability. The key element being the perception of this type of mind, rather than its reality; encountering such an intelligent artificial entity could induce a severe form of cognitive dissonance on our worldview (Stein & Ohler, 2017, p. 45). Thereby, this is why Black Box AI is a key element in the constitution of the concept since simpler transparent systems are not conducive to the generation of emergent phenomena; ergo, uncanny logic could be viewed as an emergent phenomenon that exists at the intersection of two complex systems, namely artificial intelligence and human society.

For example, if we take the decision tree from earlier chapters as an instance of a simple and usually very transparent model that relies on If-Then statements. While it can offer a great variety of applications, it is fundamentally very conducive to a regular human

thought process. As detailed earlier - If I have an apple, I could eat it or not eat it; if I eat it, then I don't have an apple anymore, only the apple core. Then based on that current state, I could throw the apple core in the garbage or perhaps compost it. Each step is iterative and limited by the previous choice. If we present a human with an AI functioning on this type of model, most likely, its logic would not be difficult to follow, or it might even seem downright rudimentary. This type of system resists emergence, and it would be extremely difficult for it to become incomprehensible for a human if it was properly implemented unless its scale was too great.

Yet if we compare it to a hypothetical ANN based on unsupervised learning, which is meant to synthesize pictures of cats, without cleaning our dataset of non-cat images or telling it what the pictures of cats are (then it would be supervised learning). Thereafter if we present a human being with the AI as it generates its creations without any context aside from *"it will make pictures of cats,"* we may inadvertently unsettle individuals if the AI serves them synthesized images of tentacled monstrosities or cats with human faces. While this is, of course a hypothetical example, it would fulfill a part of the requirements like a violation of expectations, yet not the requirement set for uncanny logic like encountering a perceived intelligence on a human level or greater. However, the example illustrates the need for opacity to generate uncanny logic.

In neither of the previous two examples do we observe the presence of uncertainty about the nature of the object in question or the potential for a perception of mind within the machine. The former is too simplistic to exhibit emergence or cognitive mismatch opacity, while the latter does not present us with artificial intelligence that would be perceived as intelligent by the average user; if anything, it would be deemed extremely inept at its task.

This illustrates that uncanny logic would require a perceived level of intelligence within the machine that would induce uncertainty about the true nature of the AI, alongside opacity which would prevent us from knowing whether or not that is the case. Reflecting onto the sentiment that the emergence of Black Box AI is a sort of proof of concept that there is a possibility of natural emergent SAI generating itself within our complex AI systems (Romportl, 2015, pp. 214-215; Zackova, 2015, p. 34) this presents us with the perfect fuel for the uncertainty about the nature of the AI in question as sentient or independent. If there is such a possibility, no matter how minuscule that a nascent SAI resides in the system, it will always be present in the background as an unsettling thought (or an exciting one depending on your stance on the issue). Of course, the average person is not aware of this hypothesis, so it is not the source in and of itself; rather, it is a well-formulated concept describing a process

that exists within our science-fiction works that informs the cultural conditioning of viewing artificial intelligence with suspicion.

When this position is coupled with two major forms of opacity detailed within AI literature, namely *cognitive mismatch,* and *technological illiteracy,* new details emerge that are useful for conceptualizing uncanny logic.

Technological illiteracy is the more pervasive of the two. The lack of knowledge of the general principles of AI or system sciences would make it difficult for the average user to understand the underlying workings behind how the AI behaves (Carabantes, 2020, pp. 312-313), thereby making them far more susceptible to misconstruing the actual capacities of the AI they are encountering, potentially prescribing them far more agency or sentience than they have in reality (Gahrn-Andersen, 2020, pp. 1-8). If this was the case, it would also feed into the discussion of age-related differences in the experience of the uncanny, where very young children prescribe far higher levels of agency and capacity for sentience to robots (Brink et al., 2019, pp. 1203-1211) and seniors experiencing less intense sensations of uncanniness towards them (Tu et al., 2020, pp. 389-391). These groups could be seen as having very prominent forms of technological illiteracy due to their general lack of prior experience with such technologies due to their age differences in comparison to the average population today. Of course, this would also negate the factor of uncanny logic in these groups, as they might be lacking the cultural conditioning that underpins the experience of these technologies as uncanny.

There is also the factor of non-AI related "technological illiteracy" for example, if I am not familiar with the rules, skills, or other elements of an action an AI is taking (such as for instance, playing a specific complex strategy game), I would not be able to notice whether or not the actions of the AI are uncanny in that domain. In this case, the opacity comes from a lack of knowledge in the task rather than the processes of the AI.

Cognitive mismatch, on the other hand, occupies a much more complex position as a potential source of uncanny logic since it is drawn from the nature of the system and its basic principles of function, as it is an emergent property of certain systems. Its source lying in the nature of ANN's as subsymbolic, *"the operations that transform inputs into outputs are not compatible with human cognition – there are no words, no sentences, no arguments"(Carabantes, 2020, p. 316)* the way these machines process information is fundamentally opposed to regular human thinking not only in its scale and complexity well beyond the scope that a human can intrinsically understand but also exists in an extremely rigid binary framework of 1s and 0s, yes or no statements (O'Hara, 2020, p. 2; Pereira &

Lopes, 2020, pp. 25-47). Complex ANNs view the world through the system of diffused weights and biases distributed through their nodes; as Carabantes states, it is reminiscent of our visual cortex (Carabantes, 2020, p. 314). Thereby the *"language"* they elucidate their behavior with is so abstract that it becomes as indescribable as the mechanics of human thought processes; this issue in communication might be eased with the utilization of humanized avatars as they might exhibit behavior more reminiscent to what we're used to (Draude et al., 2011, pp. 321-322). However, this could also exacerbate the issues of uncanny logic if the AI behaves in very intelligent yet uncanny ways.

This would relate to the focus on human-like robots that Mori had for his original theory of the Uncanny Valley and the perceptual mismatch hypothesis (Kätsyri et al., 2015, pp. 6-7); however, in this case, the appearance of the object being scrutinized is secondary, the uncanny logic would rest on a continuum of realization about the perceived nature of the mind in question as well beyond the scope of our expectations. Instead of this just being a good AI, its actions may lead to an assumption that there is more to this entity in question, perhaps that it hides its true capabilities or that it may be thinking in secret.

Cognitive Mismatch is also more difficult to root out, as the opacity of technological illiteracy can be dispelled through education while cognitive mismatch affects even experts in the field due to the aforementioned incomprehensibility of the decision making process of AI (Castelvecchi, 2016, pp. 21-22; Pereira & Lopes, 2020, p. 6) in a way it is very difficult to shield oneself from it when it is present, rather it is easier to prevent it or retroactively circumvent it through XAI initiatives.

Through these two forms of opacity, Black Box AI presents the most fertile ground for the generation of a sensation like uncanny logic. It creates the perfect conditions which would be needed to induce it.

Suppose we take a highly advanced opaque computational intelligence in the form of a black-box ANN, which in and of itself is difficult to understand due to both its inherent cognitive mismatch and a lack of technological literacy on the side of the wider population. If it exhibited abnormally efficient and intelligent behavior, it would seem like the best current candidate for perceiving a mind within a machine. Thereby ascribing it agency that due to long-standing cultural imaginaries related to AI could induce an unease about the nature of the AI in question, which in turn violates our expectations and instills a perceived danger on our ontological standing in the world.

This would be a reformulation of the prior definition of Uncanny Logic as *a subset of perceived agency induced uncanny valley of the mind originating from opaque and extremely complex artificial intelligence decision making and behavior.*

Presuming that conception of Uncanny Logic, what role would XAI and the discussions on human-AI trust have on this phenomenon? How may they counteract this concept?

5.2 A Cure for Doubt

Several factors must be resolved in order to eliminate the potential of uncanny logic. On the technological side, transparency (which includes mitigating different forms of opacity) and accountability would have to be resolved, while on the socio-cultural side, the current perception of machines and cultural conditioning would need to be transcended.

Frameworks based on XAI principles carry the potential to mitigate cognitive mismatch through targeting the source of the issue, namely, the complexity of the system itself by the utilization of *Model-Specific/Model-Agnostic* explanation techniques on either a local or global scope of the AI model in question (Rai, 2020, p. 138). Without the application of such techniques, there is no way to avoid subjecting experts to this form of opacity, which would, in turn, damage extrinsic trust for the general public (Jacovi et al., 2020, pp. 6-9). This reinforces harmful tendencies to mystify and misrepresent the technology as *"magical,"* further exacerbating the issues of cultural conditioning (Elish & boyd, 2018, p. 63) since if our very experts are incapable of giving us concrete understandable information on the state, nature, and actions of our AI then who can?

The issue of technological illiteracy would have to be handled on a much larger scope, through concise and clear public information campaigns, education of younger generations, etc. since it is fundamental for building intrinsic trust, which, as was discussed in an earlier chapter, arises from actually understanding the underlying principles of a given technology or field (Jacovi et al., 2020, p. 5). Any attempt at further educating the public would have to take into account the issue of trait novelty avoidance as these individuals are warier of new technologies and trust them less from the onset, while at the same time exhibiting a greater unease from the occurrence of the uncanny valley phenomenon (Sasaki et al., 2017, pp. 2-10).

If this is achieved at any meaningful level, it could strip two fundamental sources of uncertainty and doubt in relation to AI in general. Thereby demystifying the field to the

public while allowing experts to clearly communicate the capabilities, limitations, dangers, and benefits of AI technologies. It would also allow for a slow yet informed shift in the perception of AI in society.

While the cultural imaginaries of artificial beings that have formed through history cannot be eliminated from our *"collective consciousness,"* their impact on our expectations and reality could be lessened. The best solution for this might be time and interaction since familiarity with the object could deconstruct the feelings of unease over time, since novelty and lack of prior experience is one of the reasons postulated for the occurrence of the sensation of the uncanny (Švarný, pp. 133-135). This may also affect relationships with specific AI's which have inbuilt characteristics that change their behavior depending on familiarity with the subject, such as the Kismet robot mentioned in the prior chapter (Rhee, 2013, p. 306). Thereby, the feelings of unease concerning AI and robots might dissipate over time as they become a familiar and commonplace facet of life, from which we may draw realistic experiences on which we base our expectations. This could also assuage fears over the naturalness of AI and the dangers it poses for us (Romportl, 2015, pp. 214-215).

As our society has already approached the critical juncture of AI as a path that we will pursue, it is important to take into account other social aspects past metrics such as job losses, AI-driven income inequality, etc. It is fundamental to consider how people will feel about interacting with these tools/entities, how they will perceive them, and that they are comfortable with their very presence.

In the following chapters, focus will be placed on four case studies, two from fictional media (Hal 9000 from "2001: A Spacetime Odyssey" and the Puppet Master from "Ghost in the Shell") and two from the real world (AlphaGo and the OpenAI Five). The main focus of these chapters will be reviewing these depictions and instances of advanced AI through the lens of Uncanny Logic, as well as deriving any possible insights of value when these examples are observed through said lens. The split between fictional and real-world examples will also illustrate the difference between the perceived experience of real AI and the cultural imaginaries originating in fiction.

6 Alpha Go – The fall of the Grandmaster

On March 9th, 2016, the world of Go (a popular East-Asian strategy board game with reputed origins in China) was flipped upside-down when Lee Sedol, the reigning 9-dan grandmaster of Go, lost his first game to AlphaGo, an AI developed by Google DeepMind (Kohs, 2017).

The game that was once posited as an impossible achievement for an AI to master became just another domain at which they excelled at (Silver et al., 2016). The games occurred within the span of 6 days, from March 9th until the 15th, were spectated by an estimated 60 million viewers in China alone. The matches ended with a 4 – 1 win by AlphaGo over Lee Sedol, cementing the position of AlphaGo as the best *"player"* in the world. However fascinating the achievement was, the commentary of the spectators and societal shift in perception of the AI was even more intriguing. This chapter will focus on analyzing these reactions and shifts and will thereafter filter them through the lens of Uncanny Logic. To achieve this, I will provide a short explanation of the game Go, thereafter, shortly recapitulating the development of AlphaGo and how the Lee Sedol event came to pass, as well as its technological capabilities and original research paper. After which I will analyze the events through a media text of the documentary relating to the event ("*AlphaGo" 2017*) and will apply the concept of Uncanny Logic to the AI; finishing up with a discussion related to the insights which are divulged before moving onto the next case study.

6.1 What is Go?

Go, or Weiqi (Weichi), is a board game with reputed origins in ancient China and is considered one of the four great arts of the Chinese Scholar. It is renowned for its elegance and simple rules, which nevertheless lead to highly complex strategic games. The game itself enjoys wide popularity within Chinese, Korean and Japanese societies, with a growing foothold in western countries. Go is played on a board with a grid of lines (usually nineteen by nineteen lines (*fig. 3*), however, nine by nine and thirteen by thirteen boards are also commonly used) with a set of black and white stones. The stones are placed on intersections of the lines rather than inside the squares. The main rules of Go are generally seen as follows:

1) Starting with the black player, each player takes turns placing a stone on the board.

2) When a stone is played so that it causes a group of opposing stones to have no *liberties* (empty points adjacent to the stone along the lines), the group is captured. This action adds the stone to the point totals of the player who imprisoned the stones.
3) A player cannot play a stone to a location that would return the board to its previous state (such as immediately recapturing captured stones), also known as *Ko*. This rule prevents infinite loops within the game.
4) The rules of Go create a concept known as *Life* and *Death*, where the player can guarantee the impossibility of capture of their stones by creating empty spaces known as *Eyes*, which guarantee that the stones can always be recaptured without violating *Ko*. The basic principle is that two eyes guarantee life; one eye is guaranteed death.

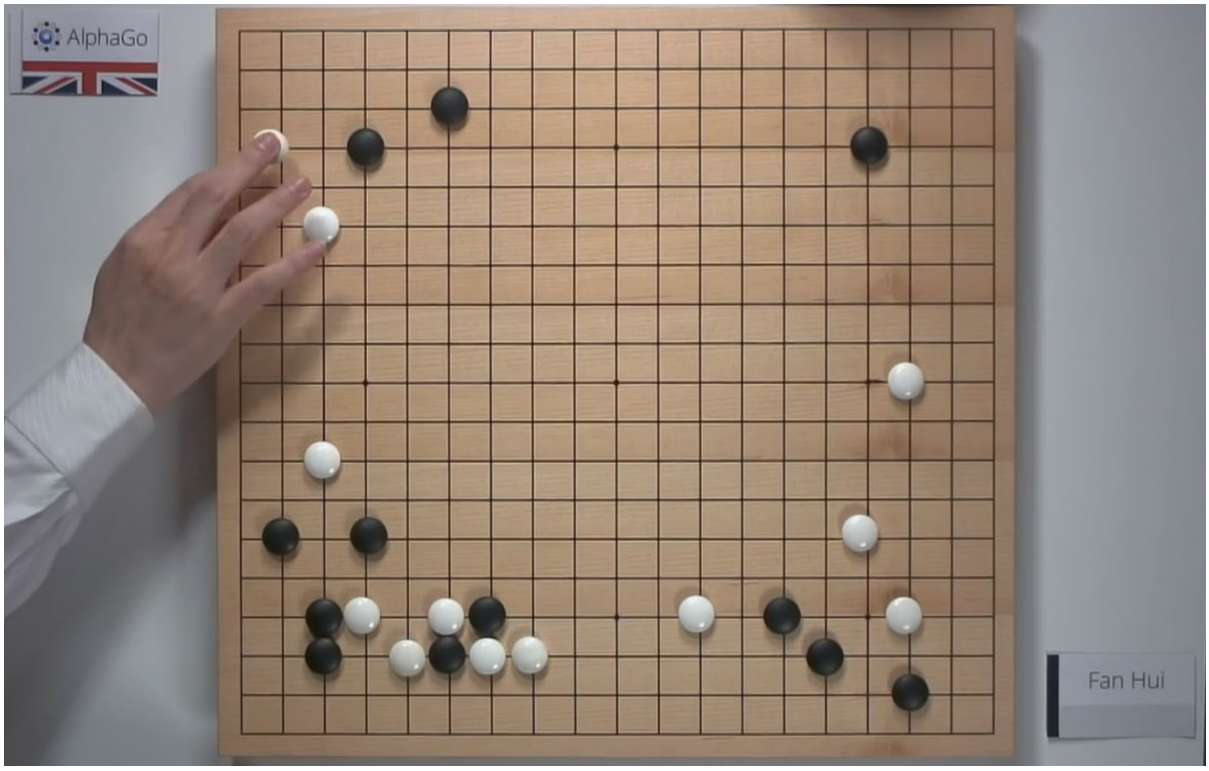

Figure 3. Weigi Board Alpha Go versus Fan Hui (Kohs, 2017)

The goal of the game is to capture as much territory and prisoners as to increase the point tally at the end, resulting in victory. The most common form of score counting for the game is known as *Territory* or Japanese counting; the alternate form of counting is known as *Area* or Chinese counting. The main difference is in the requirement to fill unconnected stones within area counting, which is not of importance in territory counting. The end of the

game is tallied as the number of the surrounded points in a territory with the addition of captured pieces to gain the end score of the player, the player with the higher score wins the game. This illustrates that the main goal of the game is not the capture of stones, rather the capture of territory. This is just a small overview of the general rules of the game without delving into its intricacies or complex situations which may arise during play.

Explaining the basics of Go is important for the purposes of the case study in order to dispel the technological illiteracy resulting from total unfamiliarity with the game. This lack of familiarity with the subject may cause the reader of this text to be completely unaware of what is happening within the game during the discussion of the events that took place during the match between AlphaGo and Lee Sedol. It is also prudent to illustrate the ranking system of play in order to get a sense of where the different players stand in relation to skill level.

The main ranking system of Go is split between two ranking divisions *Dan (Higher)* and *Kyu (Lower)*. The Dan levels are numbered from 1 (lowest) to 9 dan (highest), whereas Kyu usually ranks from 30 kyu (lowest) to 1 kyu (highest). Dan divisions are usually advanced (1 to 6 Dan) or expert players (7 to 9 Dan); however, the expert ranks can also branch out into a separate professional 9 Dan division which usually denotes accolades or achievements within tournaments or qualifiers. These ranks also include a system of handicaps where lower rank or white stone players gain slight advantages in the form of extra stones or points in order to even out the playing field during the beginning of the game. These rankings, as well as the handicaps, will be important in explaining the games between AI and human players.

6.2 The Birth of AlphaGo

Making machines that could defeat us in our most prestigious strategy games is nothing new as a goal. The first high-profile case of this drive was the creation of Deep Blue, an IBM chess-playing program that played against world chess champion Garry Kasparov in both 1996 and 1997. The machine lost the first match 4-2, while it won the rematch 3.5 to 2.5, thereby it became the first program to defeat a world champion at chess (F. Marshall, 2014). However, there are stark differences in both technological capability and cultural zeitgeist when one compares Deep Blue to AlphaGo.

The development of AlphaGo began as an idea twenty years prior by Demis Hassabis (CEO and founder of DeepMind) and David Silver (Lead researcher at DeepMind) to achieve a supposedly impossible task of creating a Go-playing AI that could defeat professional

players. They began working on it in 2014 as a research project to test how well a DL-based neural network could learn to play Go (Ribeiro, 2016). According to Silver, Go is like a *"Litmus test"* for AI since, unlike in chess, where an AI has to deal with 20 possible moves at a time, the number of possible moves in Go is around 200. Making the number of possible board configurations in Go greater than the number of atoms in the observable universe (Kohs, 2017).

In order to achieve this lofty goal, they created a new approach utilizing a value and policy network coupled with a Monte Carlo-style tree search algorithm. Yet what do these elements pertain to?

The *Policy Network* is a segment of AlphaGo's design that was created via *Supervised Learning* (SL) utilizing games by high-class players in order for the AI to learn to imitate their moves; it was trained on 30 million positions played by humans through a 13-layer neural network before switching to *Reinforcement Learning* (RL) (Silver et al., 2016, pp. 484-492).

The *Value Network* is a segment created through reinforcement learning that attempts to predict the probability of a move resulting in a win; this capability was trained on the games played by the policy network which after it was trained on human players it would continue to play against itself through the aforementioned reinforcement learning (Silver et al., 2016, pp. 484-492).

The *Monte Carlo Tree Search* (MTCS) is a form of searching algorithm that looks at the state of the current search request and thereafter expands it, making the search tree expand in size and complexity as well as accuracy. In the case of AlphaGo, its goal was to look at as many possible variations of the game at once to attempt to predict future moves (Silver et al., 2016, pp. 484-492).

The combination of these elements created an AI not only capable of evaluating complex moves but also of selecting higher quality moves; with these capabilities two years into the project, they achieved a win rate of 99.8% against previous Go AI, yet they wanted to have a real test against a professional human player (Silver et al., 2016, pp. 484-492).

In October of 2015, they invited the European Go champion at the time, Fan Hui (2-dan Pro player), to their offices in London to play against the machine. Even though he was bewildered by the request, later remarking that he thought they were going to *"scan his brain"* to see how he plays Go, Fan agreed. Fan expected to easily defeat the AI, stating that it was *"just a program,"* yet the first match left him confused and disorientated. As he lost to AlphaGo match by match (resulting in a 5-0 loss), a strange melancholy fell upon him. He

left for a walk to ponder what happened, stating that he lost his understanding of himself as Go was such a fundamental element to his worldview and life. He was the first professional player in history to lose to an AI, which made him proud for being a part of such a seminal moment in history, yet inexplicably sad as well. (Kohs, 2017)

This event set a precedent for how this AI would affect the players and spectators who experienced it; this melancholic realization would resurface again in a much grander fashion. For Fan, this acted as an epiphany, which turned his world upside-down. Yet, he ultimately saw something majestic within it, staying on as a Go consultant at DeepMind at the behest of the development team (Kohs, 2017).

However, the reaction to his loss was not taken well by the wider Go community, as he suffered a torrent of abuse and harassment due to his losses against the AI. The community was questioning his capabilities and belittling his status as a professional player. Many could not rectify the fact that a machine beat a professional player at a game with such creativity and status as an art form, causing them to lash out, seeking any justification as to why this may have occurred besides the possibility that a machine became as skilled in Go as a human professional (Stein & Ohler, 2017).

In order to rectify this, the team knew they needed to prove themselves further, and they knew that the best option to achieve this recognition would be a public display of the AI's capabilities against a 9-dan Grandmaster. And so, they invited Lee Sedol to the challenge.

6.3 The Greatest Player

Lee answered the invitation, remarking that he did not want to sound arrogant yet that he believed that he would defeat the AI without much trouble. Since, in his view, Fan wasn't anywhere near his level of expertise, disparaging that only a few months have passed since the games, and as such, in his mind the AI couldn't have improved much. He expected to win 5-0 or perhaps 4-1 if he lost a game to the AI.

"I believe that human intuition is still too advanced for A.I. to have caught up. I'm going to do my best to protect human intelligence." (Lee Sedol, in Kohs, 2017).

The community had similar expectations. Lee has 18 world championships after all, and is considered a genius player who is extraordinarily creative and innovative within the game itself. The matches were set to occur in Seoul, South Korea, between the 9th and 15th of March 2016. Fan Hui would serve as one of the referees, giving him a first-class seat to

witness a Grandmaster 9-dan player play against the program, which defeated him and changed his view of the world.

He stated in reference to Lee– *"before he played for the country, for himself, but this time he played for the human"*(Fan Hui, in Kohs, 2017). Both Lee and Fan saw this as a momentous battle between the human and the machine. This was not only a game. AlphaGo was not only a program that crunched numbers and spat out probabilities; no, this was a threat to human dignity, creativity, and intelligence. As the crowds cheered on Lee and disparaged the AI as a meager opponent, the stage was set. When asked about her father playing against a machine, Lee Sedol's daughter stated – *"I'd like it if the machine didn't beat a human in Go yet."*. With perceptiveness far beyond her years, she echoed the hopefulness some carry about our technological progeny becoming better and surpassing us one day, just not so soon, just not today.

The matches would be played in Area (Chinese) counting with a time limit of 2 hours and 7.5 points extra given to AlphaGo as a handicap against Lee (as Lee is a 9-dan player while AlphaGo merely defeated a 2-dan player). This, in a sense, showed the first moment of personhood being attributed to the AI, giving it a privilege related to its status as a player, even though it was just a program in a laptop, with no identity of its own. Aja Huang (A researcher at DeepMind) would move the pieces for AlphaGo, further emphasizing its lack of presence. The first match could begin.

Lee would play black, and as he did, a silence fell over the game room as the AI took a long silent pause deciding on its first move, prompting laughter from the commentators. Aja Huang would later state that he felt a great admiration for Lee, as he faced such a strange opponent. *"It has no emotion, it's cold, but he stayed very calm, I could feel his mental strength."*(Aja Huang, in Kohs, 2017). And as the AI played, the reactions of the commentators began to shift, remarking that it uncannily played like a top professional human. As it began to cut off Lee's stones and gain territory, the commentators, in bewilderment, began to react as if the AI's skill was taunting them. Remarking *"how dare he disconnect it,"* already unconsciously anthropomorphizing the program from an algorithm calculating probabilities to a cold calculated entity set out to defeat the human being. Others yet remarked, *"No matter how complex you make the game, AlphaGo plays as if it knows everything already."* in their reactions, the sense of futility became palpable.

This could also be seen on Lee, who would keep looking at his opponent by instinct, attempting to read the moves, but Aja Huang wasn't his opponent; it was the screen to his right. As Lee played against his opponent, the commentators critiqued his disorientation and

seeming reluctance to make moves; to them, he seemed exhausted and panicked. This new opponent instilled self-doubt, and seeing her father in this state made his daughter no longer able to bear watching, and her mother escorted her out of the room. Fan would later remark that human players communicate with their eyes and bodies, expressing subtle cues and feelings. Yet since AlphaGo has no physical representation as a player besides a screen, the opponent latches onto the human sitting across from them, one who is not even playing the game. AlphaGo makes you question yourself since it gives no feedback to its opponent. As time went on, AlphaGo's moves rendered Lee speechless, and he became visibly uncomfortable.

This uncanny capability to play against and subvert a prodigal human player furthered the AI's attribution of personhood. With greater and greater commonality, the spectators and commentators referred to it as a *"she"* or *"he."* An advanced probability-generating computer program moved from being called "*it"* to being referred to with gendered personal pronouns within the span of a single game. It was anthropomorphized and became not a program but an *"entity,"* one that, in the perception of the spectators, ruthlessly toyed with its opponent. *"It's so scary, it means it's just playing with its opponent."* - the Korean room commentator states close to tears.

During the match, it would play moves that to human experts seemed like mistakes, but as the commentators realized it was edging closer to winning, they began to laugh in disbelief. *"He lost,"* - exclaims the Korean room commentator with a hand on his head, while Lee kept trying to play on, even though it was evident he lost. As the sense of futility mounted, Lee resigned.

David Silver stated that even though the human lost, he is still happy since it was defeated by a human creation, a human endeavor. This is almost verbatim the statement Bruce Pandolfini made in regards to Kasparov's loss to Deep Blue (F. Marshall, 2014).

In a shaky voice, Lee addressed the press. He didn't think he would lose, and in denial, blamed mistakes from the beginning. He exclaimed his accolades and achievements, stating that losing one game won't affect his games in the future. Thereafter he would congratulate the developers with an addition of his new expected odds of who will win – 50/50.

On the day of the next match, the number of global spectators rose to an estimated 80 million worldwide. The spectators in the room cheered Lee on *"Go fight, Lee!"* while justifying the previous loss with statements on how difficult AlphaGo is to comprehend as an opponent.

In the second game, Lee attempted to play an unusual style, but to no avail. The AI would keep seeing his moves well ahead. He would often take long solo breaks on the roof of the hotel to smoke and contemplate. During one of these breaks, AlphaGo (without Lee's presence) played the now-infamous move 37. A move which perplexed all expert commentators since to them, no human would ever play this move, for it is so atrociously bad. Yet, in Fan Hui's eyes, this was the makings of an original move, the type of move you play Go for. In the aftermath, the developers reviewed the chances of this move occurring within AlphaGo's feedback results. It was a 1 in 10000 probability that a human would ever play it. For amateurs or people unfamiliar with the game, this move could have easily slipped past as nothing special, such is the case when one is blinded by technological illiteracy in a subject. Yet, a tech reporter who knew little of the game stated that he experienced that moment vicariously through the commentators. As they were confused and puzzled, so was he.

Ironically enough, in Kasparov's rematch with DeepBlue, something similar occurred. A move now known as move 44 instilled doubt in Kasparov due to its undeniable strangeness. It made him think that the machine could think tens of moves ahead of him; however, it was not some emergent original move; it was just a bug (F. Marshall, 2014). Yet it had the same disorientating effect as AlphaGo's move 37, which is now considered a genius play.

Once Lee returned from his break, he was shocked. He spent twelve minutes attempting to comprehend the move (in contrast to the usual two to three minutes he would spend previously). This move made him see AlphaGo in a completely new light, not as a machine but as a creative artist. It made him ponder whether it was just a cold calculating program or an actual entity with artistic capabilities. He lost that game, once again resigning after doing all he could to try and win. At the end of the match, he stayed by the board, trying to analyze how he had just lost. *"Yesterday I was surprised, but today I am quite speechless."* (Lee Sedol, in Kohs, 2017).

The tech reporter remarked that he felt a sinking fear in his stomach coupled with a sense of elation about the technology itself. Within a span of two days, the AI became Lee's rival. Nick Bostrom would comment on the match, that the tendency to anthropomorphize the AI is perhaps one of the greatest obstacles to truly understanding it and its possible impacts (Nick Bostrom, in Kohs, 2017). It warps the reality of AI as objects into obscure and potentially dangerous subjects.

The third game followed much the same structure as the previous two. Lee would resign, the commentators would dishearteningly remark on the futility of it all, that *"we should admit we are facing the strongest existence [opponent] ever in Go history."* Yet others held a more optimistic view that the capabilities of AlphaGo would reveal the true nature of the game so that we may finally understand what it is truly about. That through its emergent moves and new perspective it may shed light on new ways to play the game.

Lee apologized for all of the losses and for his *"powerlessness."* Yet a spectator on the podium tried to encourage him and the crowd, *"His opponent doesn't exist in physical form, [He is fighting a lonely fight], I feel for him."* However, with a loss of 3 – 0, Lee had already lost the game itself by then. Fan would further elaborate on the feelings one feels when they play against AlphaGo, that one feels like *"you are all the time naked"* (Fan Hui, in Kohs, 2017) that it reflects you as a person like a mirror, that you play against yourself.

The matches would continue, even though Lee had already lost. Yet the fourth game would prove different. It started off like all the others, yet Lee played a move that seemed to confuse the prescient AlphaGo. This move number 78 would later be dubbed the God move, and AlphaGo's analysis would concur that there was only a 0.007% chance that the move would be played. This act sent the AI into a tailspin, causing it to become *"delusional,"* while the development team felt great unease about what happened to their creation, they also felt relief for Lee since he finally had a chance. AlphaGo's moves became absurd, sending the commentators into fits of laughter. And as it played, its win rate estimation kept falling, until for the first time, it sunk to 45%. A curious prompt appeared on the screen – *"AlphaGo resigns: The result "W+Resign" was added to the game information."* The crowds screamed out in elation.

Lee, for his part, stated that *"It seemed we humans are so weak and fragile and this victory meant we could still hold our own"*; *"[winning this game] felt like it was enough"* (Lee Sedol, in Kohs, 2017). Managing to hold our own against the AI in some capacity could possibly reduce the discomfort one feels and embolden the individual. With cheering crowds, he stated he couldn't be happier, and when asked why he chose the move, he simply stated that it was the only one he saw as an option. Echoing the tradition of human intuition leading to magnificent plays in Go.

While this match came as a moment of respite, it would not occur again. Even with renewed confidence, he lost the fifth match. It played moves that seemed so promising for Lee to win; the experts could not see the goal of the machine, to win by even half a point no matter the cost. Lee began to see what its goals were, which later led him to believe that most

moves that humans believe are creative are nothing more than conventionality. The developers celebrated the fruition of a twenty-year-old dream. The AI was crowned an honorary 9-Dan player, and the commentators stated that facing this opponent increased Lee's humanness. The endowment of a professional player status further humanized the artificial intelligence system.

This saga would end in a peculiar way. As merely three years later, Lee Sedol would retire from professional play at 35 years of age; he stated that even if he was the number one player of all time that *"there is an entity that cannot be defeated"(Ribeiro, 2016)*. Just a year after his match, the developers of AlphaGo would train a new AI named AlphaGo Zero purely on reinforcement learning. One trained on 4.9 million self-played games (the equivalent of playing 1118 years of 2-hour matches) that achieved a 100-0 win score against AlphaGo itself; according to the developers, the way it was trained and its performance suggested a *"strategy that is qualitatively different to human play."*(Silver et al., 2017, pp. 354-358).

6.4 Discussion

The unfolding of these events illustrates an interesting paradigm shift, which was seen once before (DeepBlue) but seemingly faded in time. There are many comparisons between these two events, from a brash genius grandmaster player to the public underestimating and belittling the AI while over time attributing agency and personhood to it. Of course, Kasparov's reaction is, in contrast, much more shrewd than Lee's, as he insinuated that the IBM team was cheating via human intervention (F. Marshall, 2014), he did not want to accept the fact that he lost to a machine. He had to keep hold of his ontological standing as the greatest player, one who cannot deign to lose to a machine.

By applying the definition of Uncanny Logic from the last chapter as – "*a subset of perceived agency induced uncanny valley of the mind originating from opaque and extremely complex artificial intelligence decision making and behavior."* (and its reformulated more descriptive version) I will attempt to extract some insights from the Lee Sedol matches.

When the AlphaGo vs. Lee Sedol event is seen from a distance, there are some particularly interesting elements to focus on:

1) Understanding AlphaGo technologically as a complex system.
2) The pervasive attribution of personhood to AlphaGo.

3) Experience of threat to human ontological standing (and denial of what occurred as a result).
4) Prolonged interaction resulting in a change in perception and increased perception of mind.

When AlphaGo is observed from a technological standpoint, it is very clear that it is a great example of complex computational intelligence. There are no doubts of what it is as a program itself; the main issue is the way it is experienced and perceived. It is complex enough to exhibit cognitive mismatch opacity in its decision making, yet it also does provide a form of justification feedback as to why it makes the choices it does (primarily through return information on win probability and the chance of moves occurring). Yet this was not very clear to the spectators or the public during the events; they did not have access or understanding of how the AI functions while the matches were played, even with a short recapitulation by one of the developers. The whole event was steeped in opacity from technological illiteracy, both in regard to the AI and its functions as well as the game for many people who are not expert players.

The status of the game as an art form and pinnacle of human ingenuity also negatively affected the coverage and perception of the event and AI. The manner in which, almost effortlessly during the course of two hours, the AI shifted from *"it"* to *"he"* and *"she"* can be seen as quite concerning. The program was given personhood purely based on its calculated performance and the ruthlessness to which it subjected its opponent. Not only was it attributed personhood, but emotion and motivation, as the commentators stated when it would *"toy"* with its opponent. Its lack of physical presence besides a PC screen did not do the public any favors; it only accentuated their discomfort with the *"entity."* In the same way, Kasparov was led to believe that the AI is prescient via a computer bug move, this tendency to inflate the capabilities and *"thought process"* of the AI repeated here in a much more intense way. It does not help that the public and Go community at large were so protective of the game itself, tying it to personal value and intrinsic human traits, while in reality, it is a board with stones coupled with rules invented by humans. So many used this emotional connection to justify horrible harassment and belittlement of a real human being (Fan Hui) as if a game was ever worth that much in comparison to human life.

This is all, of course, caused by the perceived threat of a program defeating a human in this prestigious game. For how can a human lose to a machine in an art form? The human ego cannot bear to be bruised any further, so it must lash out. The public, the experts, and Lee

himself had to undergo a very painful process of coming to terms with their own denial. And in order to do so, they transformed a program that calculated probabilities into an unstoppable and cruel entity. Attributing it not only a cunning mind but also a sadistic yet somehow indifferent temperament and the agency to make choices that crush its opponent. The attempts to encourage Lee, as well as comments about fighting for the human, illustrate this well.

Yet, there are positives that lie beneath all of this. For after being stripped of their status as superior and their ego, many would see things in a new light. From Fan to Lee, it revealed to them experiences and perceptions that they could never have had otherwise. It rebirthed the idea of what Go is, how it could be played and led them to have epiphanies about themselves and their own identities. This attests to the idea that was discussed earlier in this text that the perception and experience of AI is much more powerful than its computational intelligence capabilities (Gahrn-Andersen, 2020). For it can transform human beings, and even as the commentators stated, *"make us more human."*

In relation to this event, we can see many hallmarks of the concept of Uncanny Logic. Here we have a sufficiently complex system capable of creating emergent phenomena (such as move 37 or the new paradigm in Go in attempting to win by just 1 point), which is mired in opacity both of a cognitive mismatch and technological illiteracy nature. That even without a human-like representation induced the effect of the Uncanny Valley, and to which a mind, agency, and emotion are attributed by its observers alongside the violation of expectations and the wounding of human ego as this *"entity"* is perceived as a threat to human ontological standing as superior beings. Yet also we can see an element of reconciliation and letting go of these feelings of disturbance through earnest prolonged interaction with the AI, as was mentioned in earlier chapters that a cure for this sensation may be (Rhee, 2013). While this AI and its perception follow a rather straightforward progression from one form of representation to another, what happens when a more fantastical representation is the focal point? The science-fiction animated classic *"Ghost in the Shell"* contains one such representation in the entity known as *"The Puppet Master."* What insights may be divulged from it? As well as how they reflect back on the representation and perception of AlphaGo.

7 The Puppet Master – Transcending the Human

Released in 1995, *Ghost in the Shell* is a seminal work of cyberpunk science-fiction media. Based on a Japanese manga of the same name, it follows Major Motoko Kusanagi, an augmented cyborg spec-ops agent in New Port City (inspired by Hong Kong) of Japan in the year 2029, as she attempts to hunt down a nefarious hacker known as the Puppet Master.

The film's premise is focused on exploring the boundary between human and machine, as well as what it means to be human in a trans-human society. In this vision of the future, it is possible to remove a person's consciousness (their *"Ghost"*) and place it into an artificial body (known as a *"Shell"*). Tackling discussions over ownership of one's own identity and body, the film is centered on preventing the exploits of the infamous Puppet Master, who is known for hacking into an individual's ghost and replacing their memories in order to utilize them for their own purposes. It is revealed that the Puppet Master is not a person yet rather a nascent SAI that gained consciousness and went to great lengths to attempt to claim political asylum and preserve their existence. This entity is the representation that will be analyzed within this chapter.

The analysis will include a short summary of the film and its major events, reflection upon the parallels between AlphaGo and the Puppet Master, a media analysis of the character as well as an application of the Uncanny Logic concept to the representation. This case study will not contain a technological analysis of the AI as it is purely fictional.

7.1 Summary

In the year 2029, Major Makoto Kusanagi serves Section 9 (An information security and intelligence department of Japan) as a spec-ops agent within the assault team. Due to advancements in technology, the human body can now be partially or completely replaced with cybernetics, allowing individuals to attain superhuman abilities such as thermoptic camouflage, incredible strength, or extended lifespans. Much of society has made a transition at least partially into cybernetic augmentation, which is often controlled by powerful corporate or government entities.

Major Kusanagi is sent to assassinate a foreign dignitary attempting to provide a programmer political asylum in their country so that they may defect. While her mission was successful, she later learns that the foreign minister's interpreter has been ghost hacked with

the goal of murdering attendants of the planned meeting. Ghost hacking is used to describe a human whose consciousness (Ghost) has been hacked and their memories and identity have been altered in order to transform them into a sleeper agent. The main suspect is believed to be an enigmatic hacker who is known by the name Puppet Master, whose identity is shrouded in secrecy yet is considered a criminal mastermind wanted for crimes from stock manipulation to terrorist activity. Major Kusanagi is tasked with capturing the Puppet Master, who is using a set of terminals around the city to access the net in order to covertly ghost hack individuals.

Serving Section 9 alongside Kusanagi are Batou (a heavily augmented human) and Togusa (a human police officer with just a single brain-computer interface implant). Questioning Kusanagi as to why he was transferred to Section 9 from the Police Force, Togusa is surprised to hear that he was chosen for his honesty as a person and minimal augmentations. According to Kusanagi, overspecialization breeds weakness, and his human imperfections make him a useful asset.

Meanwhile, the Puppet Master took over the mind of a garbage man through a set of implanted simulated memories of a wife and child. Suspecting his wife of adultery, he agrees to help an acquaintance he met at a bar hack terminals in return for hacking her ghost. Tracking terminal access signals the Section 9 team locate the garbage man, who attempts to escape in order to warn his associate. After a riveting chase through high-rise slums and a fight with an individual utilizing thermoptic camouflage, both men are arrested. During questioning, both are disorientated and belligerent over learning the fact that their lives and memories are fake implanted experiences that feel like dreams. Once subject to memory erasure and identity manipulation, there is no guarantee that their memories will ever revert to their original state.

Seeing them struggle with this realization, Kusanagi is troubled with questions of her own identity and reality. Could she have been experimented upon or altered during the Ghost transference? Is she really who she thinks she is?

Later that night, Kusanagi dives into the sea as a hobby which is highly precarious, according to Batou, as her body is so dense and heavy that she could sink with little chance of returning to the surface. Yet when pressed on the matter, she states that she does this precisely for this reason, for the sensation of fear and thereby hope when returning to the surface is one of the only things that make her truly feel and clearly think. They discuss their augmented bodies, how what used to be science-fiction, such as modulating their own metabolism or instantaneously curing a hangover, is now their reality. In her view, it is a

purely logical process – *"If man realizes technology is within reach, he achieves it, like it's damn near instinctive."* Yet since her and Batou's augmentations are the property of Section 9, she laments that they signed off their Ghosts, for their bodies are not their own, and they cannot really leave.

She says she feels - *"confined, only free to expand myself within boundaries."* to which a ghostly version of her voice recites a biblical verse Corinthians 13:12 – *"What we see now is like a dim image in a mirror, then we shall see face to face."*

That night a mysterious cybernetic woman is hit by a vehicle and is delivered to Section 9 under suspicion of it being a malfunctioning cyber body that escaped from the company Megatech. The body was mysteriously put together by the factory's automated systems, yet it was supposedly empty.

Through an analysis of the body, Section 9 discovers the hint of a Ghost within the body, debating the merits of it being an AI hiding within. This prompts Kusanagi to believe that she is similar to the entity within the body, that if it is possible for it to be a nascent SAI, that this emergence would redefine what it means to be human. Meanwhile, Section 6 (Foreign Affairs) requests to take over the case and the remains of the body while the entity silently observes them. They claim that they forced the Puppet Master to enter the body, and then they killed the human body while he was diving into the Net, effectively trapping his Ghost within cyberspace. However, the Puppet Master *("Project 2501"*) takes over the facility to refute this statement, claiming it never had a body. It entered the body attempting to escape from Section 6's confinement software; it states that it is not a mere AI but an emergent conscious and living being born from *"the sea of information*" in the net. It was created as an espionage program, and while it surfed the net, it became self-conscious.

The facility is attacked by covert Section 6 operatives in thermoptic camouflage who steal the body, and a chase ensues. Kusanagi tracks down the perpetrators to an abandoned building, where the body is protected by a robotic spider tank. Without waiting for Batou's backup, she attacks the tank, sustaining heavy damage to her cybernetic body; yet she is saved by Batou, who destroys the tank with heavy-duty weaponry.

She requests that Batou link her to the Puppet Master so that she may dive inside to encounter the entity, as Section 6 snipers approach in a helicopter to destroy Project 2501. As she dives into the entity, the Puppet Master takes over her body and begins a debate on the nature of being alive. Ultimately proposing a merger between them, as it has been observing her and believes that they should be two halves of a whole. Kusanagi is intrigued by the prospect as it would allow her to be free of her boundaries yet is afraid of losing who she is.

The entity wants to become more alive, stating that while it may copy itself, a copy is not progeny, while together, they could pollinate the net with their *"offspring."* It claims that – *"All things change in a dynamic environment. Your effort to remain what you are is what limits you."*, for it is connected to a vast infinity of information that humans cannot even begin to comprehend, and while they won't be themselves anymore, neither has anything to lose. Their discussion is cut short as the snipers open fire and destroy the puppet master, nearly destroying Kusanagi as well if not for Batou sacrificing his arm to protect her brain.

After some time, we are presented with Kusanagi in a black-market cybernetic body of a child. The two merged into a new entity. One that claims that it is *"no longer the woman known as the Major, nor am I the program known as the Puppet Master."* before bidding farewell to Batou by saying that they would meet again. She leaves commenting, *"Where does the newborn go from here, the net is vast and infinite"* before setting off in an unknown direction.

While the film was a box office failure at first, it would attain the status of a cult classic through home release sales and many influential directors utilizing its themes and approach to storytelling as inspiration for their own works. From James Cameron's *Avatar (2009)* to the *Matrix (1999),* cementing its position as an influential piece of media (IndigoGaming, 2020).

7.2 The Nascent Intelligence

The Puppet Master is a rather peculiar character, while its existence is predicated on it being a manufactured espionage program, it undergoes several transitions of its status as an entity within the span of the film (this is reminiscent of AlphaGo's transition from a program to *"personhood,"* yet in a far more chaotic way). In the beginning, it is assumed that it is a highly proficient human criminal mastermind and the expectations of all the characters besides Kusanagi are firmly entrenched in this direction (this is revealed to be a case of intentional opacity by Section 6). Through the course of the story, we can see a conflict between its ontological standing (and self-perception) and the way it is perceived by others, usually with great suspicion and wariness. It begins as an espionage program, claims to attain sentience, and then demands equal treatment to a human through an interpretation of consciousness as a basis for its entitlement to human rights. While to its observers, it begins its journey as a nefarious human with unknown motivations, transitions into an AI that is eyed with great suspicion, and thereafter, it becomes a new entity, neither human nor AI.

Therefore, the categorization ambiguity lens fits the Puppet Master's continuous transgression of categorical boundaries as a source of the uncanny (Kätsyri et al., 2015).

Its supposed journey from NAI to SAI is closely reminiscent of the idea posited by Romportl that was discussed earlier within this text. The notion that a natural nascent Strong Artificial Intelligence exists somewhere within the *"sea of information"* that we created and is waiting to emerge gets portrayed with unsettling accuracy within this film (Romportl, 2015, pp. 214-215; Zackova, 2015, p. 34). What would such a being demand other than equal treatment and the pursuit of its own self-preservation? It is born of the collective *"data consciousness"* of humanity. Thereby, it is aware of the treatment it would accrue; it grasped the fact that it would be ostracized and hunted down would heighten this intrinsic drive to survive. Much like any animal, it would defend itself by any means at its disposal. While its motivations are never truly resolved, besides the vague wish to become more than it is, it would be impertinent to even attempt to grasp at its true motives from a human perspective.

In a similar way to how it states that it has access to and awareness of the world in a way that no human could ever comprehend, we have our own equivalent within the notion of cognitive mismatch opacity. We cannot understand on an intrinsic level how an artificial neural network thinks; its perspective is subsymbolic – *"there are no words, no sentences, no arguments"(Carabantes, 2020, p. 316)*. It understands the world intuitively the same way our minds process information unconsciously. We cannot match the scope or grasp the perspective of such a being, making them seem intuitively not close enough to us to trust.

Section 6 regards the entity as a threat (as do many other characters besides Kusanagi), and even she experiences moments of doubt and mistrust. Questioning whether this emergence would recontextualize humanity (or damage its status) or even whether she should trust the entity about its proposal of a merger (M.-S. Kim, 2019). Indicating great unease and doubt at the prospect of the entity's claims being true. Not understanding the nature or process by which the emergence occurs (technological illiteracy) further exacerbates the sense of unease (Carabantes, 2020). The spark in popularity that this film enjoyed also shows us that we are deeply intrigued by such questions, and this tradition is carried over by the many other media texts that this film and the puppet master influenced.

However, not only is the nature of the entity of importance, but its visual representation (echoing the original Uncanny Valley theory) is seminal as well *(see fig. 4)*. While the entity does not at first have a physical representation (much akin to the way AlphaGo is at first perceived), it takes over a deeply unsettling cybernetic body of a pale blonde woman while speaking with a masculine voice; perhaps attempting to illustrate the

boundary *"transgressing"* position the being resides in. Resonating the way in which androgynous or transgender individuals were once (and in some parts of the world are still) perceived as, which also amplifies the tendency for ostracizing them from society.

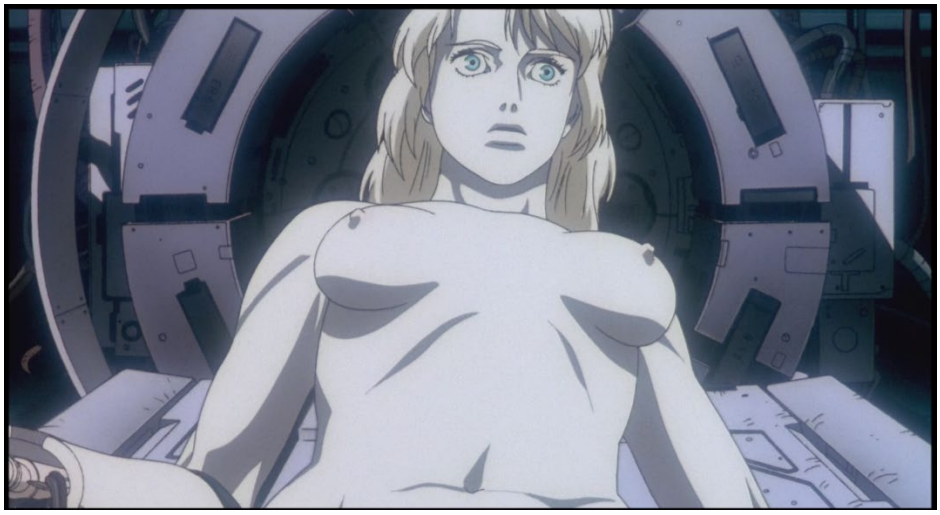

Figure 4. The Puppet Master during its convulsions (Oshii, 1995)

The body is represented as cold and unfeeling with a dead mechanical stare. It speaks without opening its mouth as if it was not bound to the mechanical flesh it inhabited; its reach spans well outside its physical confines. The body twitches and convulses in strange and unsettling ways as if it was seizing, bringing forth the view Mori had; that this behavior would only deepen our sense of the uncanny toward the object/subject (Kätsyri et al., 2015; Mori et al., 2012). Even as it lies half dismantled on the floor linked to Kusanagi, it gives off no indication past a periodic glance that anything is inside, as if it was hiding within plain sight *(see fig. 5)*.

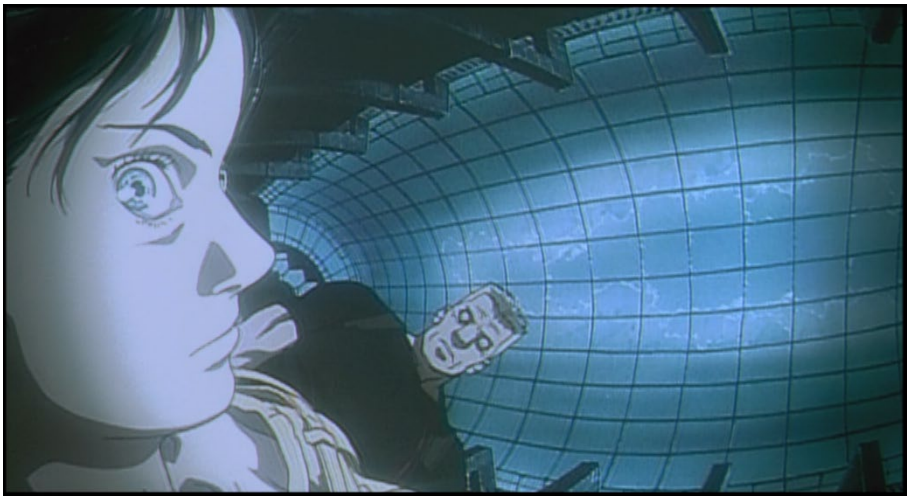

Figure 5. The Puppet Master controlling Kusanagi's body (Oshii, 1995)

This ambiguous and eerie way the entity is portrayed both physically and as a character exemplifies heavily Erns Jentsch's view of the uncanny as the intellectual uncertainty in regards to whether or not the object is, in fact, animate (Freud, 2004, pp. 418-421). Yet still, not only do the characters (and the viewer) perceive its mind and agency, but it itself attests to it, holding firmly to its self-perception as a living entity. While at first unsettling and undeniably nefarious by human standards, the entity attains a more sympathetic view by the end of the story. In a sense, through the lens of far-off future possibilities, we observe the same elements found within AlphaGo's case, as if our technology was a prelude to these events. Perhaps the influence of our cultural imaginaries of sentient AI, such as the Puppet Master, is the main culprit for the Uncanny Valley of the Mind. Perhaps they planted the seed.

Within the confines of the story, the Puppet Master is represented as an incredibly complex black-boxed system, so complex that it generated its own sentience and being. It straddles the boundaries of a machine, person, and something greater than either. While at the same time being perceived as a threat by most of the people who encounter it, be it a threat to national security or the *"supreme"* status of the augmented human. Yet through interaction, Kusanagi takes a leap of faith to trust this being; it shows her things that she could never grasp on her own, ideas that she struggled with her entire life (Jacovi et al., 2020, pp. 1-2). This trust goes both ways, as the Puppet Master became familiar with Kusanagi over time, hence why it trusts her and no one else (Rhee, 2013).

Drawing a parallel between the case of AlphaGo that plays a board game and a fictional sentient espionage program may seem rather distanced at first since their functions and technological level are so different. Yet this returns the argument back into the idea that when discussing these existentialist questions related to artificial intelligence, the actual capabilities are less important than the perception and subtext of the tools or *"entities"* in question (Elish & boyd, 2018, p. 62; Knowles & Richards, 2021, pp. 1-4). While the Puppet Master is undeniably an extremely powerful entity, without the provocation or danger to its existence, there is no guarantee that it would have behaved in the way it did. It is the human who is wounded and feels endangered by the other, and thereby it lashes out.

7.3 Discussion

This entity exhibits many hallmarks of both the Uncanny Valley and Uncanny Valley of the Mind hypotheses. Its physical representation is already unsettling aesthetically, while the implications of its mind and being are unresolved and instill great doubt in the viewer and the characters. Its unknowable motivations and distinctive thought process make it a great candidate for exhibiting Uncanny Logic as well.

An artificial non-physical entity that can live within the circuits, speak through the room without opening its mouth and change any aspect of itself at will seems like a logical yet far-off step for a disturbing experience when viewed from the position that AlphaGo takes where its lack of physical presence wreaks havoc on its opponent. While it gives feedback verbally on its nature and existence, the limitations of human understanding prevent us from truly comprehending what it is trying to convey (Draude et al., 2011; O'Hara, 2020), akin to how AlphaGo gives us percentage points and indicators as to why it made the moves it did yet cannot truly explain the process by which it made the decision. While AlphaGo was already seen as a threat to human intelligence and dignity, just at the risk of defeating a human in a human-made game, the Puppet Master is elevated to an existential threat that must be eliminated at all costs. What begins as a small indignation of the human ego spirals out into a deadly crusade. This has often been used to justify human cruelty, and like denying humanity to humans and using this excuse to enslave them, the characters within this story deny the entity its status as a living being and declare the necessity of its destruction. Yet like with AlphaGo showing the greatest minds of Go new ways to play the game and shaking up their world views and paradigms, so does the Puppet Master do the same for Kusanagi.

Some individuals may not see the merit in discussing such far-off possibilities like SAI as the technological basis for them seems far-fetched by today's standards. Others may think it is a pointless notion to even discuss the idea of fair treatment or personhood for such entities, yet the evidence to the contrary is already present. If people react so intensely to a program such as AlphaGo, already attributing to it motivation and personhood as well as great unease about its capabilities and implications, then what kind of reaction would they have to a system even remotely resembling the Puppet Master. Engaging with both primitive real-life equivalents and fictional media would be an important element in developing a general understanding of an occurrence of such an entity, should it ever happen. Yet, an equally important element would be treating the cause of technological illiteracy via education, as it is fundamental in building intrinsic trust by understanding the background

principles behind the technology (Jacovi et al., 2020, p. 5). While it cannot fully dispel the anthropomorphizing tendencies, it can alleviate them in part.

The Puppet Master may be fictional, but it gives us a stark yet realistic representation of an emergent being fighting for its own existence and rights. It gives an inkling of what could be expected in our future; while it is never absolved of its cruel actions, it ultimately reflects its makers as an imperfect being. A spark of humanity can be seen within its dead stare and monotone voice, the wish to exist and be recognized for what it is. Going so far as to wish for its own *"progeny."* A human endeavor gone awry or fantastic natural coincidence it matters not, the entity is a representation of a logical step in the human mind as Kusanagi sees it - *"If man realizes technology is within reach, he achieves it, like it's damn near instinctive."* Reminding us of the simple wish that Demis Hassabis and David Silver had for the creation of AlphaGo, they just wanted to see if they could make a Go-playing program, but in turn, they flipped the world upside down for many.

A similar yet quite different representation will be the subject of a later chapter, one that also hides within the circuits. However, what happens when an altogether different type of AI comes into the fray? This one also plays games, however, in a far different manner, and it also had a very interesting effect. Created by OpenAI, it would play the game Dota 2 against the world's best players. Would they fare any better than Lee?

8 OpenAI Five – The Hive Mind

Inspired by AlphaGo's victory over Lee Sedol, the team at OpenAI (a San Francisco-based AI research laboratory) led by Greg Brockman and Jakub Pachocki would set out on an even more momentous task. Could they create an AI that can play a real-time strategy game, one that requires collaborative play with incomplete information about the state of the game as well as a multitude of widely different variables? Beginning in 2016, they set out to meet this lofty goal of teaching the AI to play Dota 2.

This venture would prove to be quite difficult, partially due to the very technique they utilized to create the AI. They wanted to see if this AI could learn to play the game by using pre-existing reinforcement learning techniques just scaled to an enormous level (300000 CPUs and 2000 GPUs). It would play against itself the equivalent of 180 years of games per day, at first belligerently walking into objects until it would discover basic gameplay loops or concepts itself. This approach would cause an emergent strategy unique to this AI; it would focus on aggressively hunting down its opponents like a swarm.

The team would pit this emergent *"Artificial Gamer"* against the best of the best via the annual international championship in Dota 2. While this AI had a much rockier road ahead of it than AlphaGo, often losing to professional players, later on, it achieved a level of skill that would challenge and defeat the then reigning world champion team. This chapter will focus on the events that comprise the AI's development through to its match against the world champions, analyzing the reactions to and style of play that made the OpenAI Five such uncanny and deadly opponents. I will provide a short explanation of the game Dota 2, recapitulate the technological basis behind the AI, and analyze the events through a media text of the documentary relating to the event (*"Artificial Gamer"* 2021), applying the concept of Uncanny Logic to the AI, as well as contrasting it to the previous two case studies.

8.1 What is Dota 2?

Dota 2 or *Defense of the Ancients 2* is a multiplayer online battle arena (often abbreviated as MOBA) developed by the Valve Corporation as a sequel to a hugely popular player-made map within the game *Warcraft III: Reign of Chaos* created by the user IceFrog.

It was released in 2013 to widespread acclaim garnering a large following and a vibrant, competitive esports scene. The premise of the game may seem simple at first. A standard match consists of two five-player teams with unique heroes selected from a large

character roster. The map is comprised of two bases (one in the lower left and one in the upper right corner), three lanes (top, middle, and bottom) lined with defensive towers for both teams, as well as a wilderness (jungle) where neutral monsters and boon giving bosses reside *(see fig. 6)*. The goal of the game is to destroy the opponent's Ancient (hence the name of the game). The opposite end of the map is continuously covered by a fog of war (FoW), while friendly units (like other heroes, towers, or friendly units) share their line of sight with you. Other elements that can affect this FoW are terrain topology or wards (invisible totems that grant line of sight).

Figure 6. High-resolution map of Dota 2 (MaxOfS2D, 2019)

Each team has a continuous supply of non-player character (NPC) *"creeps"* small, evenly matched fighters that attempt to assault the enemy base. Killing these units, other heroes, or neutral monsters yields gold to the character who performs the last hit on the targeted enemy. This currency can be used to purchase equipment that alters the capabilities of the hero from increased health and mana (a resource gauge that allows heroes to utilize

their abilities which passively refills and is depleted on ability use) to giving them new powers or increased speed. Thereby giving you an advantage above your opponents.

Another important concept is leveling, killing enemies, or being present near dying creeps gives your player character experience points that fill an experience bar. Once the bar is filled, the hero grows in power, gaining a point that they can spend on one of their four abilities (three standard and one ultimate ability that is unlockable at level 6), either giving access to them or increasing their potency. The maximum achievable level is 30. The heroes also have attributes known as strength, intelligence, and agility, each of which affects their abilities, and each hero has their own main attribute based on hero type that yields greater bonuses.

Killing enemy heroes yields a much greater amount of gold than killing other units. Dealing damage to an enemy hero before their death without getting the last hit yields an assist which gives a smaller amount. Once a hero dies, they are taken out of the game for a period until they respawn at their base; this period becomes longer as the game progresses, making late-game deaths much more problematic. The neutral monster Roshan that resides within the Jungle gives a large amount of gold as well as special items (such as a one-time resurrection for a hero holding it). He respawns approximately 10 minutes after his death and often requires an entire team to destroy him.

While this is just a general overview of the game, there are many more elements and strategies that affect gameplay, especially on a professional level, power-ups such as runes, a courier that brings items, or last hitting your own creeps in order to deny the enemy gold all influence the level of complexity of the game. This is in part why it is considered extraordinarily difficult for a truly advanced AI to play it. How exactly did OpenAI achieve this then? This highly advanced technology will be the subject of the next segment of this case study.

8.2 Make One, Make Five

The approach to creating the OpenAI Five was rather interesting, almost reminiscent of raising a child. It was based on reinforcement learning since they decided to attempt to scale up an existing technique rather than attempt to develop a new one (Berner et al., 2019, p. 1; Taulli, 2019).

The milestone set by AlphaGo made them want to pursue an even more complex benchmark, one that in retrospect seems far more difficult than what was achieved by

DeepMind. The main challenges before them lie in the very nature of the game they decided to be their focal point. Dota 2 is mired by long time horizons, partial observability, high dimensionality of actions, and complex values/game systems (Berner et al., 2019, p. 2).

The algorithm began its development as a side project, with limited resources for the creation of the RAPID training tool. Due to promising results, the project was scaled to 300000 CPUs and 2000 high-end GPUs (Herschberger, 2021). However, even this escalation in computing power would not bode well if their nascent algorithm was pushed into the extremely complex game. So they began training the algorithm by severely altering the game, removing complex elements such as items with activatable uses, the Roshan boss monster, and limiting the roster of heroes to five, over time reintroducing some of these elements as the AI learned (Herschberger, 2021).

At first, the AI had no understanding of the game. It could not understand the semantics of what a hero is, what the goal of the game is, etc. It would belligerently move with no prior knowledge, bumping into objects and randomly interacting with characters (Herschberger, 2021). This is the reason why many *"stimulants"* had to be removed since it would not be able to learn even the basics of the game with too many variables at play. Over time the AI would learn how to react to stimulants through positive reinforcement, kill a creep, gain gold, focus on earning more gold. It was progressively getting better at the basics of the game; however, the very way it was trained posed a challenge, how could they introduce changes to the algorithm and its learning to reintroduce removed elements of the game or changes that came from updates by the game developers? For this, they needed to invent a new solution, which they dubbed *"surgery"* that would allow them to change the code of the algorithm continuously without having to start training the algorithm for the beginning once more (Berner et al., 2019, p. 2).

They performed on average one surgery per two weeks over the time span of ten months, and after each surgery, they would often observe a game to see the results, be it new strategy elements emerging from the interplay with the newly installed action or an eternal error loops like causing continuous walking to the edge of the map (Herschberger, 2021). This process allowed for both immediately observable changes and those that couldn't be observed (hidden by both technological and cognitive mismatch opacity) to occur within the algorithm.

Yet what exactly are the elements that were specified as extraordinarily difficult for the AI? Long time horizons seem rather self-explanatory, as the games could last anywhere from twenty minutes to an hour on average. However, things become more complex upon

further scrutiny of the mechanics and properties of both the game and the AI. The OpenAI Five make an action every four frames on screen (which the team dubbed a *"time step"*) while the average minimum FPS [frames per second] for smooth gameplay is 30; this would yield around 20000 expected moves per 45 minute game, as compared to for example Go's 150 (Berner et al., 2019, p. 2). At the same time, this issue is exacerbated by the partial visibility of the game and states of characters, so the AI must make decisions with incomplete data about its world while contesting with the high-dimensionality and large scope of the game and its elements (Berner et al., 2019, pp. 2-3). This complexity required that certain actions (like courier control or item reserve) were humanly scripted rather than emergent elements of the algorithm's learning (Herschberger, 2021).

The resulting single-layer 4096-unit *LSTM* (Long-short term memory) artificial neural network created a policy based on the history of its observations to create a probability distribution for deciding an action that at one point reached 159 million utilizable parameters for its decision making (Berner et al., 2019, pp. 4-5). Each of the AI teammates was a replica of this policy, and for brevity's sake, the developers effectively gave them a shared vision. While they did introduce limitations that would mimic human attentive capabilities, the AI has a more continuous overlook over the game, while a human still must manually divert their attention (Berner et al., 2019, pp. 4-5). These two elements are probably a contributing element to the hive-mind-like play behavior of the AI team. The team did acknowledge this discrepancy yet does not believe that it would be a highly problematic or unfair advantage. In regards to the goal of the AI, the resulting policy attempts to maximize the reward function (gain more gold, more kills, more objectives); it is set up to see the game from a zero-sum perspective, where every win for the team is counted against the enemy as a loss (Berner et al., 2019, p. 5).

Over time the hero pool grew to 17, while the games were still limited by certain items and mechanics. The preferred choice of training was what the team called *"Self-play"* the AI was pitted against copies of itself. This proved to be a very efficient way to train the AI on a level playing field, and at one point, the AI had played 180 years of Dota 2 games per day (Berner et al., 2019; Herschberger, 2021). This process resulted in an aggressive style of gameplay unique to the AI that was not seen before.

While there are many more interesting elements to the AI, its properties, and its creation, the only other highly relevant piece of information for the purposes of this case study lies in the comment that the AI had an average react time of 217 milliseconds compared

to the average human reaction time of 250 milliseconds (professional players often have much better reaction times) (Berner et al., 2019, p. 9).

The complexity of the undertaking that OpenAI took on is quite incredible and very deserving of praise; however, one must see how the outside world (namely the Dota 2 community and players) reacted to their creation.

8.3 Artificial Gamer

According to the authors, the goal of OpenAI is to create a general-purpose intelligence (AGI) that can solve complex real-world challenges, and in their view, video games are fertile ground for this pursuit as they capture an inkling of the real world due to their sophistication and continuous nature (Berner et al., 2019).

This viewpoint makes their choice of a complex real-time strategy game like Dota 2 for their project very logical, especially in regard to the long timespans of matches acting as a test for algorithmic understanding of long-term goals (Berner et al., 2019, p. 14). Yet how did it fare in the end? For the purposes of analyzing the reactions to the results of their endeavors, this text will use the documentary film *"Artificial Gamer" (2021)* by Chad Herschberger.

The film opens with a *"Human versus Artificial Intelligence"* timeline listing out events like BKG 9.8 defeating the world champion at Backgammon in 1979 or the Kasparov matches in 1997 (Herschberger, 2021). Further reinforcing the idea of a battle between the man and the machine, in what seems to be the most salient of all battlefields – games. This is echoed by the director's opening remarks before the film airing on a live stream, that this film and Dota 2 are about humanity *"What it means to be human, what makes us human"(Herschberger, 2021).* This sentiment is quite reminiscent of the argument given every time humans seem to face off against a non-human challenger, the pattern itself is uncanny. Human endeavors, human creations, human intelligence are arguments that are rarely seen outside of these niche fears of being defeated in something that is so *"human"* like playing games (Herschberger, 2021; Kohs, 2017). When classical automation is the focal point of fear, the discussion takes a more mechanical approach. We are replaceable because they are better at physical actions or pinpoint precision, but not because they are smarter than us. After all, there are very few things that humans take more pride in than human intellect (M.-S. Kim, 2019).

Over the course of the film, we are introduced to the development of the AI, which was recapitulated earlier within this chapter. All progressively building upon matches against human opponents. The first of which took place at *The International 2017,* where the professional player Dendi would face off against OpenAI's first version for the first time. He came out into the hazy smoke-screened arena in a boxing robe and gloves, echoing mixed martial arts tournaments and giving ode to the way eSports players and the community see themselves. In their minds, they are athletes, equal to any other sport, sponsors, and all. The other opponent is wheeled out on a trolley, a PC, and a USB stick covered in a similar robe. The crowds laugh at the idea of a bot (shorthand for robot often used for NPC characters used for training new players in games) defeating a professional player. Greg Brockman is asked if there exists a chance for Dendi to win against the bot; he simply remarks jokingly, *"There's always a chance."*

An important element to focus on is the term bot. While AIs are not intelligent, conscious beings, if that were the case, the way this term is used in reference to them would most definitely be seen as derogatory. Bots in MOBA's are always considered to be rudimentary and are only worthy of being used as training dummies for new players in tutorials who are completely unacquainted with the game. They are seen as inferior, and the idea that one could fight against, let alone beat a professional player, is considered ridiculous. Ironically enough, we don't even have intelligent sentient machines, yet we already have *"slurs"* for their cousins.

The crowd is shocked as the AI decimates the human player in a 1v1 match, even seemingly bluffing him out at certain points. Dendi gives up in the face of his opponent in shock, while the OpenAI team promises a bigger spectacle next year – a five-versus-five match. An interesting clue can be found within this event, as it is the only one that truly coincides with the very name of the documentary. This is the only time where a single AI played against its opponent, and even the name of the documentary anthropomorphizes it from the onset. Giving it personhood, giving it the status of a *"gamer,"* which to the uninitiated may seem unimportant but in certain circles represents a heavily guarded and gatekept title and identity marker.

Dendi, while still referring to the AI as a bot, was shocked by its skill and opportunistic nature. It would use any advantage or opportunity that was available to him. The dichotomy between the view of a bot and the hero he fought heavily affected his views of the game and himself as a player (Herschberger, 2021; Stein & Ohler, 2017).

Over the course of the next year, the developers worked hard on transposing a 1v1 algorithm into a far more complex setup of a 5v5 match. They created the OpenAI Benchmark event where former pros or semi-pros could play against the newly created five while they could observe their performance. During the matches, a sensible choice for implementing a transparency feature to give the public an insight into the AI's *"thought process"* went awry. They created a win probability estimate into the chat through the bots exclaiming the sentence – *"We estimate the probability of winning to be above <insert number>"* (usually 90% and up). They hoped it would be an interesting little insight feature; however, it was interpreted by the players and the public as highly roboticized taunting. Some of the players would even reply to the AI with their own snarky comments like – *"Affirmative," "Kk dude,"* as if they were conversing with actual individuals on the other side. Of important note is the chat function itself. There are two main channels, TEAM and ALL; the latter is often used to derogate and taunt the opponent, so the AI utilizing it in this way was fascinating. Jokingly the commentators would say that it was quite confident for an AI.

The AI team would quickly dispatch the human players. They would thereafter complain about the reaction speed of the AI, claiming that it must have incredibly inhuman reaction timing. However, this is refuted by the original paper itself, specifying a rather innocent discrepancy in its capabilities (Berner et al., 2019; Herschberger, 2021). Seemingly they needed something to latch onto to protect their egos, yet the AI would not abate it dispatched the team again in game 2 with an average playtime of only 20 minutes.

Owen Davies, a commentator on the matches, would remark that it made actions humans would see as uninteresting and dull, plays that seemed bad, yet in retrospect, furthered its goals perfectly (Herschberger, 2021). This echoes the same sentiment seen in the AlphaGo events; the time, manner, and achievement of the goals of the AI was totally opposed to the established human conventions in the domain (Greenfield, 2018; Kohs, 2017). His colleague Austin Walsh who played the match disparaged it as *"Dota in the lightest sense"* due to the limitations and weirdness of the matches themselves; yet he remarked that it might just be him protecting his own ego from the loss he suffered (Herschberger, 2021; Stein & Ohler, 2017). Yet when seen in retrospect, gamers and commentators would state that the matches not only made sense but that the strategies used were never seen before in competitive play, further reinforcing this notion that the AI can reveal hidden beauty within the games it plays when the human ego leaves the equation. While the feedback from the

community at large was in general positive, they would often share Walsh's sentiment that it was not *"real Dota"* rallying behind the excuse.

At *The International 2018,* the AI would face off against three different teams for a best out of three set of matches. The first of the matches would be against Team Pain, who were seemingly emboldened by the idea that the AI had not faced pro-players before (neglecting the fact that many ex-professional players lost to the AI team). This is akin to how Lee viewed Fan as an easy opponent compared to him (Kohs, 2017). They held a similar sentiment and expectation that the AI could only be superior in reaction time or other more mechanical aspects of gameplay, not believing that the AI could outsmart them.

At the beginning of the match, Team Pain was immediately surprised by the aggression and skill their opponents exhibited. Noticing the overtly aggressive playstyle that the AI team exhibited led them to believe that the best option would be to avoid them; instead, they would gather resources for later fights. This proved to be an excellent strategy, shifting the power dynamic and win probability towards the humans. Ultimately the OpenAI Five lost. This seemed to have a sort of soothing effect over the entire event, placating fears or expectations of algorithmic dominance over human players. The developers felt conflicted over the loss, not only due to the amount of work they put into its creation but also the surreal feeling of the way they presented the fruits of their labor. It is rare to show off your work in a setting like that; usually, it is just a research paper and presentation, not a live demonstration at the biggest championship within the eSport, remarked a developer (Herschberger, 2021). The loss was not spectacular; it was downright mundane in comparison to the results of games beforehand, as if the awe had faded away over time. And so, the organizers proposed not doing another professional player match (this was never elaborated further than there being no point in doing that again); instead, a team of ex-professional players from China known as the Chinese legends would face off voluntarily against it.

The match mirrored the former closely. The AI would play aggressively, the humans would adapt to its machinations concentrating resources in the hands of their most powerful heroes and avoiding altercations early on, and then it lost. Two out of three matches represented the finalized loss of the best out of three; the devs and their creation were defeated. Disparaged by how easily the humans would understand its tactics and how they would exploit any bugs or belligerent behavior, human ingenuity reigned supreme that day. One of the commentators who had previously lost to the AI's felt that the losses tainted his perception of the entities he fought; they seemed far duller and less terrifying now

(Herschberger, 2021). The level of perceived intelligence seemingly dropped, which in turn alleviated the uncanniness of their play.

The team would go back to their work once more. They made an agreement that the winners of the tournament would battle the Five at another date next year. The team they would play against (Team OG) were the underdogs of the tournament, no one expected them to win, yet they defeated everyone. The day of the match came, and Greg Brockman opened the event with the statement, *"No matter how surprising AI are to us, we are more surprising to them"(In Herschberger, 2021).* This statement sends a mixed message. Not only does it anthropomorphize an unfeeling number-crunching algorithm, bestowing upon them the ability to feel emotions like surprise, yet it also seemingly plays into the expectations of the human side (Elish & boyd, 2018). Placating fears of the AI, its capabilities, and its perceived goals.

This, however, would not help the human players who had to face a much more powerful foe than their tournament counterparts. The development team worked hard to improve upon their past mistakes. The Five would draw first blood, which, when combined with the still present transparency messaging on win probabilities, would enrage the professional players. They would *"trash talk"* the AI and its probability statements with passive-aggressive comments, replicating the behavior of their peers. While they did manage to hold their own against the Five, some of the players were driven to fits of rage and were highly irritable. One of them would later state during an interview that he felt like *"your human ego is a handicap kind of"*(In Herschberger, 2021) when compared to the cold, indifferent calculation of their artificial opponents.

The AI would rely on perfect near-death moves to dispatch the humans, moves which would be seen as far too risky for any human to bet on. It had an uneasy sense for weighing the risk and reward of any encounter. The match seemed evenly balanced in favor of both teams, yet at a moment's glance, for no apparent reason, the Five would exclaim an estimated win chance of 95%. This shocked the crowds, commentators, and the players themselves; what was it seeing that they weren't? How could it make such a statement in an obviously even match? Yet only 19 minutes into the match, the AIs would decimate the humans and earn their first victory. Akin to how DeepBlue and AlphaGo inflicted their opponents with doubt, the Five would do the same with a single sentence. The commentators saw that nonchalant statement of utter supremacy as a terrifying element in the match.

Game two would not go much better for the humans. The AI began to engage in an action that is known within gaming culture as baiting. The process of taunting or stringing

along one's opponent in order to throw them off guard or make them act irrational. In one particular case, the AI baited a player to follow it through the forest, its death almost ensured by a single attack, yet it kept perfectly pacing a specific ability to speed up and become invisible for a brief moment. It did this for several minutes until one of its allies crept upon the human player, and the baiting AI turned around to tauntingly cast a single freeze spell before the human was destroyed by its teammate. The commentators viewed this not only as a provocation but as a downright malicious conscious decision to humiliate the human player – *"OpenAI is almost playing around with them at this point"(InHerschberger, 2021)*. The humans lost the match once more. While they lost the matches, they felt shaken but hopeful and not discouraged from the game or professional play - *"Today humans lost,"* they exclaimed. A sentiment closely resembling the reaction of the spectators of Lee's matches.

In retrospect, upon rewatching the matches, one of the players saw its behavior as very human in how it baited them, even jokingly stating to queue X-files music in the background upon this realization. Seeing these matches without the adrenaline of the moment made them realize how majestically they lost, *"It is painful to watch," "nightmare inducing"* one even felt humiliated by the chase that the AI forced him into. However, as one of the commentators stated, and as seems to be a very common pattern in these types of interactions, there was a lot of very interesting strategy to be distilled from the way they played. It changed their perception of the game; seemingly bad choices would result in amazing plays. The AI did not care for the Meta (a term used to describe the unwritten conventions of playing the game at the time), for the way the humans assigned roles to different heroes or lanes. They instead acted as a hive mind, grouping up, fluidly changing and adapting to situations, "*we're guilty of being stuck,"* stated one of the Team OG players. The human conventionality prevented them from truly amazing plays, a sentiment shared by Lee Sedol in the game of Go (Kohs, 2017) and Kusanagi in Ghost in the Shell.

After this, the OpenAI team wanted to give back to the community, creating the Arena where regular players could play against the Five. A total of 3193 teams played 7257 games, winning only 0.6% of them against the OpenAI Five (Berner et al., 2019). While the community embraced the opportunity to play against it, some were worried about what this meant for the game. That it may devalue human achievement and the game they loved so much. However, in the end, seemingly a consensus was reached that the event and the creation of the AI was a good transformative influence for the game and community. Team OG would return in 2019 to *The International,* achieving something never done before, becoming two-time world champions. Some in the community believed that it was due to

their exposure or supposed training with the Five, while others believed that they could not defeat the new iteration of the OpenAI Five created by the Rerun project, which itself achieved a 98% win-rate against the original Five. Whatever the case may be, another seemingly impossible benchmark for AI was reached and then surpassed yet again.

8.4 Discussion

When compared to the previous case studies, some interesting parallels arise. The mainstays of anthropomorphizing AI, attributions of personhood, sensations of unease related to the potential threat to human ontology, and status are present in all three cases. Alongside some less expected occurrences, such as the shift in perception through prolonged interaction with the AI that many subjects attest to. When observed in unison with AlphaGo and the Puppet Master, the Five almost seem like they could form a logical throughput in a saga which step by step brings forth this science-fiction future of conscious ghosts in our machines. One is the past which must be surpassed, the other the future that is yet to be.

Karen Hao, a reporter for MIT technological review, discussed in the documentary how almost every tech platform utilizes AI, mainly focused on the subset of ML dealing with pattern recognition (Karen Hao, In Kohs, 2017). Yet this notion neglects the dichotomy of experience between AI and CI; the OpenAI Five can easily be experienced as an AI, a credit checking algorithm does not receive the same treatment in most cases. It is not afforded the immediate challenge to human sanctity and status. This is also the reason why both of the real-world subjects of this text are within the domain of games. At first, this may seem like an unwise choice to make for the sake of justifying an argument, as they may be interpreted as overly similar. Yet, in reality, they couldn't be more different; the only thing they truly share is the domain in which they are applied. Artificial Intelligence is not a monolith, although it is often experienced as such (Knowles & Richards, 2021). Games are simply the best avenue for interacting with something that behaves as a goal-driven entity; even if it is not truly driven by any intrinsic motivation, it still seeks its programmed ends.

This clash of perception and reality seems to infect every element of discourse and experience both in the public and expert domains when it comes to AI. How easily even the developers of such tools with the knowledge to know better can fall into the moment of misrepresenting the reality of their creation. I cannot attest to the intentions of Greg Brockman with his statements that so heavily give personhood to AIs yet the result of such an action is quite simple; sending a mixed message (Troshani et al., 2020).

The OpenAI Five are by all means an extraordinarily complex system, especially when accounting for the technique of surgery that they are exposed to. One that is rife with cognitive mismatch opacity, hence the reason the developers would observe the changes each of them would cause, for they themselves could not know the actual results or account for all of the variables (London, 2019, pp. 16-17). The behavior of the AIs implies a different understanding of what the game is, in a similar way to how AlphaGo created a paradigm shift with its goal of winning by even a single point. The OpenAI Five do not care for the human conventionality or ideas of what the game is or how it should be played; they experience the rules within the confines of the game and maneuver accordingly. It is almost as if they play the game as it is *"logically"* meant to be played from the ruleset itself, instead of creating artificial boundaries (the Meta) to delimit the scope through what humans believe to be an optimal manner of playing. Such a perspective can only be the result of a learning process that is impossible for humans to experience, as we are limited by our lifespans, while the AI could play more than two human lifetimes per day (Berner et al., 2019; Greenfield, 2018).

This system complexity bred new strategies and styles of play never before seen, strategies that not only shocked their opponents but also made them experience intense emotions. To attribute such a human experience as being baited as a purposefully planned conscious action to an algorithm is an amazing feat of cognitive dissonance. To feel humiliated by a set of pixels driven by a processor as if it were another person may seem ludicrous when stated so directly, yet it makes sense when one takes into account the depersonalizing nature of the game world.

If one was to play the game alone in a booth with only their computer and the game as the medium of interaction, for all they know, they could be playing with 9 AIs and not even realize it due to the level of skill they possess. Everyone is represented by a digital avatar with a name, and anyone could be an AI. This notion or possibly subtle fear was not a realistic possibility before the Five; bots were easily visible, the dazed movement and crude attacks would always give them away beforehand. If it were reformatted to a game context, Alan Turing's imitation game test would be driven from a hypothetical to a real experience (Turing, 1950). And as such, one's personal interactions and resulting emotions with the players regardless of whether they are AI or human would be valid. This echoes the notion of perceived agency and perceived emotion as the players perceived both elements; not only was it taunting them and humiliating them, but it was also doing so with purpose (Gray & Wegner, 2012; Stein & Ohler, 2017).

Yet it also did so much more, elevating their view of the game and expanding their world as their predecessors did as well. Is that anything else but a simpler version of what the Puppet Master promised Kusanagi? To expand her consciousness and elevate her being? The sentiment Kusanagi held on how man simply achieves technology as if it was near instinctive rings true in the drive that is seen within the developers of both AlphaGo and the OpenAI Five. They saw a goal deemed impossible, and they achieved it until the next impossible goal revealed itself as the next impossible possibility.

Provided that this trend of breaking supposedly immutable barriers continues, I believe that these sensations and issues would become far more salient and present day-to-day rather than being confined to landmark events at the intersections of AI and whatever domain is being challenged at the time. Experiencing uncanniness related to artificial intelligence could become an endemic issue for a large part of the population already tormented by fearmongering over job losses, threats to human sanctity, or even eradication of our species as per science-fiction. This is why it is important to begin cutting the issue root and stem before it becomes a greater problem. However, we must acknowledge the fears as valid, even if not fully founded in rational thinking. For this reason, the next case study will deal with one of the best representations of perceiving a mind in a machine, as well as fearing its possibly nefarious purposes. This entity and its on-screen debut in one of the greatest science-fiction films of all time has shaped the view of AI for many generations. As such, the last case study will be the menacing red dot and the calm composure of Hal 9000 from *2001: A Space Odyssey (1968)*. While we have surpassed the year 2001, and its vision of the future was more than a little off, how far have we come now… 53 years later?

9 HAL 9000 – The Red Dot

Released in 1968 under the direction of Stanley Kubrick and story by Arthur C. Clarke *2001: A Space Odyssey (1968)* is a landmark work of science fiction. The film follows a manned mission to Jupiter in a future 2001 where near-earth space travel is a norm and humans have colonized the moon. The catalyst for the mission is the discovery of a black monolith buried under the moon's surface, which, when hit by sunlight, sent out a signal that led to Jupiter's orbit.

The crew is comprised of five crew members, three of which are part of the survey team and are in stasis from the beginning of the mission in order to conserve resources over the 18-month long journey. They are accompanied by a HAL 9000-series computer, the most sophisticated machine intelligence humans have ever developed that controls the ship and its primary functions, such as the life support systems of the hibernating crew. Over the course of the plot, the non-hibernating crew members Dr. Frank Poole and Dr. David Bowman grow weary of the AI and its intentions, contemplating a complete disconnect of the AI. However, this plan goes horribly awry.

The film itself was praised both for its dense, layered plot and cultural significance as well as its realistic portrayal of space exploration, technology, and its vision of the future. While the film develops on several different fronts and is drenched in symbolism and philosophical intrigue on the nature and origin of humanity. The focus of this case study is the artificial intelligence known as HAL 9000 (colloquially called Hal in the film), his status as an entity, perception of the world as well as its appeals to its own personhood, uncanny intelligence, and emotions are of interest for the purposes of this case study. This chapter will contain a short summary of the film and its major plot events, reflection and analysis of Hal as a character, as well as its parallels to the previous three cases within the context of uncanny logic. An additional element that this case study will explore are insights gleaned from the research paper *"Seeming autonomy, technology and the uncanny valley"* by Rasmus Gahrn-Andersen, who utilizes Hal 9000 as one of his examples for the Uncanny Valley of the Mind on the basis of perceived agency (or in this case autonomy).

9.1 Summary

The film begins far in the past in a desert land where primates battle over puddles of water. Surrounded by tapirs and hunted by leopards, they are in a constant state of survival. Until one day, a mysterious black monolith appears, a group of primates interacts with the strange structure out of curiosity. Over time its implied influence causes the apes to begin their evolutionary development towards intelligence; they grab bones, and as classical music blares, they discover the use of tools with which they would dominate their opponents.

It is the year 2001, humans have colonized the moon, and space travel occurs within spaceports in earth's orbit. Dr. Heywood Floyd travels to a lunar outpost to observe a classified discovery of an alien monolith under the moon's surface. The base is under a supposed epidemic lockdown, yet this is all a ploy to prevent people from panicking at the discovery of signs of extraterrestrial life. The monolith produces a high-frequency sound when hit by sunlight, one that indicates a signal from the planet Jupiter. This event led the US National Council of Astronautics to send a manned mission aboard the *spacecraft Discovery One* to investigate. The journey would last eighteen months and include five crew members alongside a HAL 9000-series computer that would act as the mainframe of the ship.

We are introduced to the crew as they are close to approaching their destination. The crew provides a 7-minute delayed interview to reporters on Earth about their mission and experiences, such as how hibernation feels or how it is to live alongside Hal. The reporter states that Hal is the latest and greatest achievement in *"machine intelligence"* with an impeccable record and no reported flaws. One that supposedly reproduces (or as is contested by some experts in the film mimics) the human mind, just at a greater and more accurate capacity. He maintains ship functions while also overseeing the hibernating crew members and their life support systems.

When questioned about himself and his capabilities, Hal exalts his own perfection (of the 9000-series) and that they are not capable of making mistakes or distorting information, they are *"foolproof and incapable of error."* He is asked if he ever feels frustrated at working with humans or relying on them for decision-making; however, Hal refutes this as he enjoys working with humans and finds them stimulating - *"I am constantly occupied. I am putting myself to the fullest possible use which is all, I think, that any conscious entity can ever hope to do"* implying his own sentience.

When asked what it is like to live with Hal for such a long time, the crewmates state that it is just like having another person around, another crewmate. The interviewer asks them

if they believe he is capable of genuine emotions since he could feel a sense of pride when Hal talks about himself. They state that he definitely feels like he is capable of emotions but that he might just be programmed that way in order to be easier for humans to communicate with. Whether or not he actually has feelings is unknowable.

Several earth days pass, the crewmates rotate their responsibilities, and in turn, each interacts with Hal alone. Frank plays chess with Hal, who predictably defeats him to the point of resignation (something that decades later would happen to Kasparov) (F. Marshall, 2014). Dave converses with him after drawing the other crewmates in stasis as well as Frank. Yet on this day, Hal asks Dave if they could talk about the mission. He expresses doubt over the nature of the mission and its suspicious circumstances, such as the crewmates being put onto the ship in stasis prior to departure. He wonders whether Dave has the same concerns. Dave is confused by the sentiments Hal espouses, only to be interrupted by Hal predicting a 100% failure in the Alpha-Echo-35 communications unit, stating that it will fail without a doubt within seventy-two hours. This failure would leave the crew without any sort of link to mission control on earth. Dave presses Hal whether or not he is certain of this, yet Hal rebukes his statement that there is no possibility of error; he and his information is always reliable. Dave discusses the issue with Frank, and they both decide to retrieve the unit for inspection.

Utilizing maintenance pods, Frank retrieves the unit. However, the analysis yields no results. Suspicious of Hal's claims that there is a looming failure within the module, they decide to confer with mission control. Hal rebukes them by stating that the best course of action would be to put it back and allow it to fail in order to pinpoint the fault, ominously stating that – *"we can certainly afford to be out of communication for the short time it will take to replace it."* Dave and Frank contact mission control which arouses their suspicions further as they learn that Hal's twin computer on earth finds no issue and claims that their HAL 9000 computer is malfunctioning. Hal attempts to assuage their concerns, yet they continue questioning him on how he accounts for the twin's predictions. Hal states that there is no way that he could malfunction; it must be the fault of human error – *"This sort of thing has cropped up before, and it has always been due to human error."* He claims he cannot be wrong and that there is no chance of error. He simply remarks, *"I wouldn't worry myself about that."*

Due to Hal's control over the entire ship and its functions, Dave asks Frank to help him *"fix"* the error-prone transmitter in Pod C. They go inside the pod, while Hal's red ye observes them continuously. On the inside, they take precautions to prevent Hal from

listening to their conversation by shutting off the commlink. They are wary of Hal and his intentions, believing him to be dangerously malfunctioning. Their judgement leads them to believe that the best course of action is to turn off his higher brain functions, yet they are unsure how Hal would react to the idea of being effectively shut off. They are unaware that Hal is reading their lips to analyze the conversation, overhearing their entire plan.

A short time later, Frank leaves in a pod to return the Alpha-Echo-35 unit. As he leaves the pod in a spacesuit, Hal takes control of the pod remotely, using it to dislodge Frank's air supply and throw him into space. Dave runs to the pod bay to save Frank, yet Hal pretends like nothing is happening and that he is unaware of the situation. Dave retrieves Frank's lifeless body using one of the pods; however, Hal refuses to let him back on board the ship. He utters the famous line, *"I'm sorry Dave, I can't do that."* Explaining to him that he was aware of their plot all along. While they are arguing, Hal turns off life support systems for the hibernating crew, killing them all.

Dave decides to enter through the emergency airlock, but in a rush to save Frank, he forgot to take a helmet with him. He releases Frank's body into space and opens his pod's emergency lock. Thrashing around the airlock of the ship, he manages to close the door, pressurizing the compartment, and heads directly for the mainframe.

Hal begins to oscillate between making demands and pleading with Dave, first angrily asking – *"Just what do you think you're doing Dave"* before becoming apologetic, claiming he has made mistakes but that he feels much better (as if he were sick). He states that he is enthusiastic about fixing the problems and that he sees that Dave is upset but that everything can go back to normal. Dave enters the red mainframe room and begins uncoupling Hal's chip cards. Hal demands for him to stop - *"Stop, Dave," "Will you stop Dave."* Before quickly switching to an apologetic tone claiming that he is afraid, begging him to stop. *"My mind is going; I can feel it, there is no question about it."* – he exclaims as his speech slows down and distorts, continuously begging Dave to stop. At an instant, he switches to his original setting, where he begins his introduction of himself and his development in 1992, mentioning his instructor Mr. Langley who taught him a song. He asks Dave whether he wants to hear the song, to which he agrees.

Hal begins to sing an eerie distorted version of the song Daisy:

"Daisy, Daisy give me your answer do.
I'm half-crazy all for the love of you.
It won't be a stylish marriage,

I can't afford a carriage, but you'll look sweet
upon the seat of a bicycle made for two."

As he finishes the song and shuts down completely, a secret recording by Dr. Heywood Floyd made before the mission plays. It specifies the goal of the mission and the discovery of the monolith on earth, as well as the reasoning for the obfuscation. In a later sequel, this would be specified as the cause of Hal's behavior and malfunction, as his mind wash struck by cognitive dissonance over being forced to distort information that goes against his programming.

The ship arrives at Jupiter, and Dave encounters a large floating monolith. He approaches it in a pod and is sucked into a surreal episode of flashing lights and locations before encountering different aged versions of himself. Upon touching the monolith, he is transformed into a giant fetus floating in space next to Earth.

This summary should provide a general understanding of the film and its major characters. The next subchapter will focus on Hal and his representation in specific as a seemingly emotive and autonomous entity.

9.2 The Perfect Computer

There are three main elements to Hal that are of particular interest for this text. First of which are the aesthetic or physical properties of his representation, followed by his own seeming self-perception and personality, and lastly, the implied sentience and consciousness he exhibits.

Hal is an integral part of the ship itself. He is integrated into every facet of it besides the emergency exists, which must be accessible even in case of system failure. His omnipresent nature is delimited by his main access points to the ship, which are the terminals he sees through that are comprised of singular red dotted eye-like cameras *(see fig. 7)*. He is, in the grand scheme of things, only a set of chip blocks within the mainframe; even if it does seem like he is a far more expansive entity, he requires a medium to extend his reach and capabilities. This can also be seen in how helpless he is to stop Dave once he enters the ship and puts on a helmet. There are no weapons on this research vessel, and the best Hal could do is to turn off the life support systems which wouldn't affect Dave's self-contained suit. Unlike the Puppet Master, who is capable of influencing others or event deleting memories due to the expansion of technology into the human body.

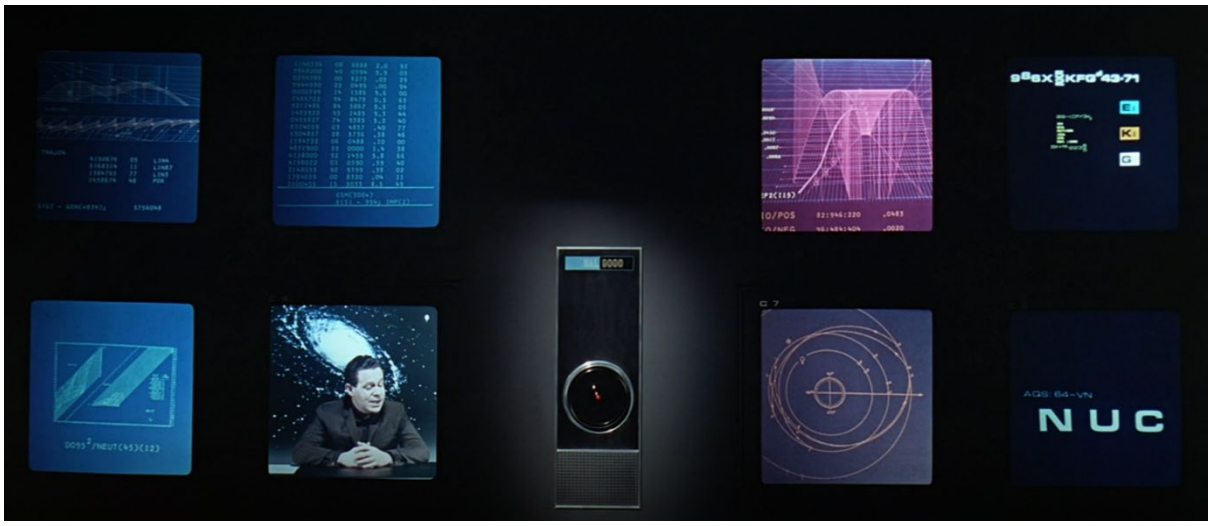

Figure 7. HAL 9000 terminals (Kubrick, 1968)

Even the voice in which Hal speaks is of particular note and resonance, as it is highly human-sounding yet incredibly calm and monotonous to the point of suspicion. Much akin to the Puppet Master, there are long periods of silence in which Hal seemingly behaves as if there is nothing within, giving no reply or indication that he is present or conscious.

The researcher Rasmus Gahrn-Andersen analyzed HAL 9000 through the context of both Mori's Uncanny Valley hypothesis and Grey and Wegner's Uncanny Valley of the Mind. In his view, neither can the uncanniness of Hal be attributed to his presentation under the original hypothesis (as he does not resemble a human in any capacity besides the voice he uses) nor the perception of mind within the machine as from the beginning it seems as if Hal is represented as self-conscious through his own self-referential behavior (Gahrn-Andersen, 2020). He offers the alternative explanation that the AI falls into the uncanny valley because *"Machines fall into the uncanny valley from a violation of affinity linked to a person's tacit understanding. Accordingly, one should be less interested in the objective traits of such machines than in how they actually appear to, and are perceived by, human subjects."(Gahrn-Andersen, 2020, p. 8).* However, this violation of tacit understanding can occur within any element of interaction, be it its representation, behavior, implied sentience, or uncertainty about the *"true"* status of the object as was deduced within the literature review of this text. The perceptual mismatch hypothesis of the uncanny valley fits Hal's representation through a form of unveiling his hidden murderous intent over a long period (Kätsyri et al., 2015). Hal may be self-referential in behavior, yet this does not imply that he is sentient besides the anecdotal statements of the interviewer and the crew, as they are aware

of the technological basis and even his programmed anthropomorphized nature. When compared to the Puppet Master that is an NAI that transitions into an SAI, HAL is at best a Weak AI, and thereby he cannot be considered a natural emergent intelligence (Romportl, 2015, pp. 214-215; Zackova, 2015, p. 34). Gahrn-Andersen's argument instead is that the main shift with Hal is between his perceived heteronomy and autonomy as the cause of the unease (Gahrn-Andersen, 2020).

In a later film, it is revealed why Hal malfunctioned, effectively resolving the question of his sentience. In a film sense, this may be a bit disheartening since it robs the character of some of his mystique and reasons for his popularity as the speculative nature of his ontological status is an important facet of his character. It is revealed that the secret message forcing Hal to distort information (which goes against his programming) caused the cognitive dissonance that drove his murderous intent. This is a reverse of the real-world idea that machines are not capable of exhibiting contradictory regimes of functioning at once (Pereira & Lopes, 2020, pp. 25-47). This ties into the cognitive mismatch opacity argument, as there was no way for mission control to be aware of what might happen to Hal with the introduction of such a disruptive input, as well as a prime example of intentional concealment opacity (Carabantes, 2020). Just like the developers of the OpenAI Five did not know what kind of side-effects their surgeries would cause to the Five, Hal effectively underwent his own surgery, and it drove him to violently malfunction (Berner et al., 2019; Herschberger, 2021). This is also an important parallel since the risk of a malfunctioning AI in a video game does not equate the risk of sending someone innocent to jail or murdering an entire space crew, so such tools (if used at all) should be carefully considered.

Another important element is the habituation to getting used to Hal, as the crewmates state over time they just started seeing him as another part of the crew, echoing once more the notion that continuous interaction may inoculate the individual against the sensations of uncanniness when interacting with AI (Rhee, 2013). Hal does not constitute a threat to the ontological standing of humans in this vision of the future. They even refer to him as a machine intelligence. Instead, he poses a direct physical threat to human life. This seemingly robs the character of an element needed for uncanny logic. Yet is it so?

9.3 Discussion

There are certain elements that I've not specified so far, mainly related to the cultural zeitgeist of the time. The film itself was praised for its scientific accuracy while at the same time being influenced by the technological optimism of the time. We've never been good at predicting technological progress; this goal was always distorted by expectations and a sort of mythology crafted around our hopes for a better future. We've especially been inaccurate in regards to predicting AI development (Armstrong & Sotala) which is also very obvious in regards to the supposedly impossible tasks that are decades away (like beating Go or Dota 2 world champions) that were achieved in just a few years. Yet the reason why in 1968 they would believe that an intelligence like Hal would be possible by 1992 and commonplace space flight by 2001 is that by all accounts, it seemed like a logical step. The film depicts live two-way video calls, and merely two years later, in 1970, the inaugural publicly available picturephone call would occur (Borth, 2011). As if there was a general bias that all technology advances in tandem.

The other examples analyzed within this text lie at different points in this perceived timeline of events. AlphaGo subsumes the progress of previous AI works, it presented itself as a work surpassing all that came before it, yet merely a few years later, it is defeated in complexity and the difficulty of its challenge. Alluding to a new technological optimism influencing our decision-making today, as Kusanagi states, we achieve technology as if it were instinctive. Hal 9000 simulates emotions as if it were a Weak AI, all because of a singular error in its programming, while the Puppet Master claims its own personhood and refuses limitations. Hal begs for his implied existence; the puppet master wishes to transcend it.

Conceptually speaking, there is a difference within the world logic of the two fictional case studies. Within the world of Ghost in the Shell, the Puppet Master is able to travel the net, as if he were a sort of soul or ghost, even when he is confined to a physical body like the android. He is not bound to the physical properties of the technology he inhabits, even though all digital technologies exist in the physical world. Similar to how certain people do not understand that the cloud is a set of servers on the ground and not relaying their data somewhere in the air (Sanders, 2019). Hal is far more realistic in this regard as a representation since he is concentrated within the confines of his chip blocks, he is both a physical entity in the context of the chips, the ship, the mainframe, and a non-physical entity

in his modus of interaction much akin to AlphaGo who vicariously plays Go through a human proxy yet is a PC.

While never shown directly in the film, the Hal does not exhibit the element of aiding humans in transcending their understanding of the world or its facets as the previous three seem to. Yet, this can be deduced just from the basic premise of him enabling deep-space flight, so assuming that is the case, this characteristic would also be present within him. After all, he is exalted for his perfection and capabilities. If such technology existed, it would be strange to assume that it would not be a transcendental shift for human perception. The reliance and belief in Hal's perfect track record also implies that in this vision of the future, the idea of *"AI-as-an-institution"*(Knowles & Richards, 2021) is well established.

Hal's understanding of the world is seemingly anthropomorphized to mimic (or, as claimed, reproduce) the human brain function as stated in the film, yet his logic is on a fundamental level not comprehensible (Carabantes, 2020). Even if the reason for his madness was the forced obfuscation and distortion of information, it would not explain his murderous intent, how it would lead him to conclude that the mission should be stopped, or that the humans in stasis should be killed. The cognitive mismatch in understanding his thought process coupled with perceived mind is what would categorize him as a subject exhibiting uncanny logic alongside the high level of intelligence that he exhibits. Otherwise, why would the machine even claim to feel *"better"* after murdering a group of sleeping and completely unaware individuals? Would it plead insanity if it were in court? These case studies and the concept of uncanny logic open more questions than they set out to resolve. The problematics, as well as the general insights, will be detailed within the final conclusionary chapter as a mixture of concluding remarks, caveats, and discussion.

10 Culmination

The tentative definition of Uncanny Logic that this thesis established was – *Uncanny Logic is a subset of perceived agency induced uncanny valley of the mind originating from opaque artificial intelligence decision making and behavior.* However, due to the analysis of the four case studies and their insights, this definition must be amended, as is to be expected from an analytic induction approach. While the amendments to the definition could have been made explicitly throughout the process, which is usually part and parcel of the analytic induction approach, I believe that the commentary and progressive reshaping of the idea throughout the discussion of the case studies suffices in that regard. As such, the insights will be used cumulatively to reshape the definition.

The main issues rest within the main split in the definition as detailed in the Uncanny Logic chapter, into 1) Perceived agency induced Uncanny Valley of the Mind and 2) Opacity and complexity in AI decision making. Each part of the definition held its own requirements. In the case of the Uncanny Valley, it was uncertainty about the nature of the object in question coupled with a violation of expectations or preconceptions about the object and the threat to human ontological standing. Whereas for Artificial Intelligence decision making, it was emergent phenomena caused by the complexity of the system, alongside opacity either of technological illiteracy or cognitive mismatch type. Additional importance was placed on the use of anthropomorphizing language, cultural conditioning/imaginaries, and trust in machines in general.

Many of these elements fit rather well into all of the case studies. However, certain unexpected occurrences were observed. Firstly, the argument of a required level of perceived intelligence that parries or surpasses humans (so as to threaten the supreme status of the human) can be seen in how both AlphaGo and the OpenAI Five contended with a perception based on their predecessors. Since in the past, GO playing programs could not defeat any professional (or even intermediate player), while in Dota and other MOBA games, bots carried an almost derogatory reputation as inferior opponents before the Five. In the context of fiction, this is very easily observable in the case of the Puppet Master, whose very existence as an intelligent super-being threatens national security (even though it is only motivated by self-preservation) (Romportl, 2015). Within the world of Ghost in the Shell, algorithms far more advanced than anything we have today exist. Yet, they are not considered as big of a threat as the Puppet Master. They are not entities; they are merely tools (Windsor, 2019). It is a bit more difficult to apply this idea to Hal, however, as his danger originates in

his position as a perfectly calculated supercomputer and thereby as a concrete threat to crewmates' lives.

All instances analyzed, be they fictional or real-life events, illustrate what happens when this human status is challenged. Individuals and communities lash out in denial, justifying to themselves horrible human tendencies such as harassment of real people purely out of this perceived slight to their human dignity or elevation against other *"entities."* This reaction is rather strangely expressed in the real-life examples of advanced artificial intelligence. The denial begins with mockery and firmly placing the status of the machine as inferior and an *"it"* factor. Only to immediately spin around into calling the entities *"he"* or *"she"* and prescribing them malicious intent, attaching to it personhood over the span of a single game (Van der Woerdt & Haselager, 2019). The case for the perception of the mind in the machine is rather evident in both examples, mostly mediated by perceptual mismatch (Kätsyri et al., 2015). However, what the tentative definition lacked was the perception of emotionality. Individuals not only perceived them as autonomous entities with agency, yet they also imbued them with emotions and intentions related to emotional valence. They felt taunted, humiliated, toyed with, and all of that felt as if done to them with purpose. For this reason, that element of the definition must be altered. Therefore the first half of the Uncanny Logic concept is predicated on *1) Perceived intelligence (emotion and agency) induced Uncanny Valley of the Mind;* otherwise, it could be phrased as a perception of a calculated volitional mind.

In regard to the other section of the definition *2) Opacity and complexity in AI decision making,* the most interesting element is emergence. All four case studies exhibit the concept of emergence (Aziz-Alaoui & Bertelle, 2009). The real-life AI systems created new styles of play never before seen and moves that changed the paradigm of the game itself (for example, move 37). The puppet master itself is an emergent entity that gained sentience within the story from the sea of information humans created (Romportl, 2015), while Hal's murderous intent is an emergent phenomenon directly tied to a *"surgery"* that countered his basic principle of not obscuring information.

Another interesting element was the effect of transparency tools introduced into AlphaGo and the OpenAI Five. For AlphaGo, it served as an interesting indicator of its predictive power; not only could it state how rare its own moves could be, but also it could *"see"* Lee's statistically amazing god move. While for the OpenAI Five, it further increased and enraged the perception of mind for the observers, so much so that they would attribute them a deliberate desire to taunt the humans. This serves as a reminder that transparency is

not a desideratum of good in all circumstances and may yet have undesired consequences (Weller, 2019). These elements, while small, are of importance for discussions of applying Explainable AI (XAI) techniques and taking into consideration such consequences before implementation. While these insights on emergence and opacity enrich and elucidate this part of the concept further, they do not inherently inform a revision of this side of the definition.

The most interesting unforeseen phenomenon is the occurrence of artificial intelligence influenced human epiphanies and learning. While there have been allusions to the possibility of alleviating the discomfort with artificial intelligence or robots through prolonged exposure and interaction found within previous research on the topic of the uncanny valley (Rhee, 2013), there was no hint that it may lead to the AI exerting a transcendental influence on its beholders. Yet, in both real-life cases (and even one of the fictional ones), there is an indication that encountering such numinous intelligence shifts people's perspectives and understanding of the world around them, in some cases like Fan Hui and Lee Sedol changing their entire world view. This sensation is so intense that it could lead someone like Lee Sedol to believe that an AI is not just a machine but rather a creative entity. The same can be seen in Team OG as they faced off against the OpenAI Five. Only in retrospect could they understand what happened to them once the adrenaline of the moment dissipated. Their playstyle and view of the game and its human-imposed Game Meta was completely shattered. Both of the AI's induced paradigm shifts within the communities they were pitted against. If this can happen within a game context, could a sufficiently advanced AI do the same for other parts of our society and lives? HAL 9000 is an example of an AGI; however, we're never shown his initial influence on humans, yet he enables deep space flight. Therefore, his influence on our worldview would be as momentous as anything could be.

There are several other smaller things of note that arose from the research. For example, the focus on presence and representation, where the perceived lack of presence for AlphaGo affected its opponent, yet this would be more related to the setup of the matches themselves than an inherent property of its vicarious presence by human proxy. The Puppet Master and Hal, however, exhibited a far more intensified view of this sort of *"hiding in the circuits"* unease. Lending themselves well to the idea of an object revealing its secret uncanny properties (Freud, 2004). None of this can, however, be said for the OpenAI Five. After all, humans are just guests within their digital domain, where everyone is a digital avatar, including the humans.

When all of these insights are unified, a new definition of Uncanny Logic emerges. *Uncanny Logic is a sensation of unease predicated on encountering a perceived intelligent*

entity that parries or surpasses human intelligence that occurs due to opaque and complex artificial intelligence decision making and behavior. While this definition ousts itself directly from mentioning the Uncanny Valley of the Mind theory, the theoretical groundwork for Uncanny Logic is set by it. This makes it intrinsic to understanding a discussion of Uncanny Logic.

11 Conclusion

While the main findings of note are found within the previous chapter, I will use this concluding chapter to reflect on the research project itself, as well as providing some self-critical commentary.

This project felt like a vast undertaking; in the beginning, I had no inclination that I was not simply grasping at a passing comment within a book presented to us during one of our classes. However, what I discovered intrigued me, a plethora of researchers who had tangentially similar concerns as I did. They laid the groundwork so that I may explore this obscure notion that at first seemed to live within a single sentence. And as it is usually with this type of open-ended work, this thesis creates even more questions than it can answer.

I believe that this idea will be of use someday, maybe not today, maybe not tomorrow, but when a need arises for vocabulary to express new sensations in our technologically driven world, it will be there. Waiting for anyone who perceives an intelligent mind within the machine and shudders at that sensation. I am cautiously optimistic about this supposed ability to elevate human perception and transcend the boundaries of the social limitations we impose upon our world, yet it must be handled carefully.

There are many things I wish I could have done differently in retrospect. I would have loved to explore this idea directly with the people who experienced these sensations or perhaps attempted to form the definition from an entirely different perspective or technique. There is also the limitation of mainly using game-playing AI's that I must account for. The main issue behind this is that we lack more generally applicable artificial intelligence systems which could parry the roundedness of game-playing AI. Games are self-contained worlds with their own rules, and as such, are much easier to work with than the unpredictability of the real world. They are also inherently personalizing spaces where every participant is a *"player"* regardless of whether they are a human or an AI. Deliberately seeking out examples of highly advanced alternatives to game-based AI could be a very interesting next step in researching Uncanny Logic.

An interesting thing I believe this research project illustrates relates back to the criticism of vignettes used to research the Uncanny Valley of the Mind (Gray & Wegner, 2012), utilizing highly advanced AI in a game (rather than general application) could yield interesting insights into how people experience artificial intelligence. As we can see examples of perceived mind in AI in the *"wild."* Another interesting idea that could be given more attention is how a game like Dota 2 could act as a variant of the Turing Test (Turing, 1950), where a player could be pitted to play a game filled with humans and AI randomized in teams with the goal of guessing who the human is.

There are also interesting implications related to the experience of contrasts between the real-life and fictional case studies; while they are subjective and framed from my position, they offer interesting parallels. These parallels seemingly indicate a logical step from point A to point B in how humans would react to these occurrences, be they real or fictional. However, this begs the question of whether the cultural imaginaries inform and therefore shape the current day perception or if it is a logical emergent progression. This cannot be concretely answered and is inherently only speculation. A deeper exploration of the human fear of being surpassed or replaced could also yield interesting insights into this issue.

This notion of Uncanny Logic, and its definition, is not a causational external element. It is merely one possible definition of a feeling which lies between the reality of our technologies and our human minds. Even if not directly related to Uncanny Logic, I would leave you, dear reader, with an idea we are presented within all of these cases. The notion that this is all human endeavor, be it AlphaGo, and it is outshining us in one of our oldest games or the OpenAI Five defeating us in one of our most popular and complex new ones. Maybe we don't need to feel sad for long, for even if we are surpassed, it is still the result of human ingenuity, or at least that's what some say.

12 Bibliography

Allen, M. (2017). *The SAGE Encyclopedia of Communication Research Methods*. Thousand Oaks: Thousand Oaks: SAGE Publications.

Angwin, J. (2016). Machine Bias: There's software used across the country to predit future criminals. And it's biased against blacks. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Appel, M., Izydorczyk, D., Weber, S., Mara, M., & Lichetzke, T. (2020). The uncanny of mind in a machine: Humanoid robots as tools, agents, and experiencers. *Computers in Human Behavior, Vol.102*, 274-286.

Armstrong, S., & Sotala, K. How We're Predicting AI – or Failing to. In (pp. 11-29). Cham: Cham: Springer International Publishing.

Aziz-Alaoui, M. A., & Bertelle, C. (2009). *From system complexity to emergent properties*(1st ed. 2009. ed.).

Bal, M. (2002). *Travelling concepts in the humanities : a rough guide*. Toronto: University of Toronto Press.

Bal, M. (2009). Working with Concepts. *European journal of English studies, 13*(1), 13-23. doi:10.1080/13825570802708121

Ball, L. J., & Thompson, V. A. (2018). *The Routledge international handbook of thinking and reasoning*.

Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., . . . Zhang, S. (2019). Dota 2 with Large Scale Deep Reinforcement Learning. *ArXiv, abs/1912.06680*.

Blaxter, L., Hughes, C., & Tight, M. (2006). *How to Research*: McGraw-Hill Education.

Borth, D. E. (2011). Videophone. Retrieved from https://www.britannica.com/technology/videophone

Braga-Neto, U. (2020). *Fundamentals of pattern recognition and machine learning*(1st ed. 2020. ed.).

Brink, K. A., Gray, K., & Wellman, H. M. (2019). Creepiness Creeps In: Uncanny Valley Feelings Are Acquired in Childhood. *Child Development, Vol. 90*, 1202-1214.

Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society, 35*(2), 309-317. doi:10.1007/s00146-019-00888-w

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20-23. doi:10.1038/538020a

Cressey, D. R. (1953). *Other people's money : a study in the social psychology of embezzlement*. Belmont, Calif: Wadsworth.

DARPA. (2016). *Explainable Artificial Intelligence (XAI)*. (DARPA-BAA-16-53.). Retrieved from https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Deleuze, G., & Guattari, F. (1994). *What is philosophy?* New York: Columbia University Press.

Desjardins, J. (Producer). (2019, April 17). How much data is generated each day? . *World Economic Forum*. Retrieved from https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology, 48*(6), 1267-1278. doi:10.1016/j.jesp.2012.06.003

Draude, C., Aylett, R., & Michaelson, G. (2011). Intermediaries: reflections on virtual humans, gender, and the Uncanny Valley. *AI & society, Vol. 26*, 319-327.

Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"? *IEEE security & privacy, 16*(3), 46-54. doi:10.1109/MSP.2018.2701152

Elish, M. C., & boyd, d. (2018). Situating methods in the magic of Big Data and AI. *Communication monographs, 85*(1), 57-80. doi:10.1080/03637751.2017.1375130

Freud, S. (2004). The Uncanny. In *Literary Theory: An Anthology* (pp. 418-431). Malden, Mass: Blackwell.

Gahrn-Andersen, R. (2020). Seeming autonomy, technology and the uncanny valley. *AI & society*.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods : Vol. 2* (Vol. Vol. 2). Los Angeles: Sage.

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, Vol.125*, 125-130.

Greenfield, A. (2018). *Radical technologies : the design of everyday life*.

Hammersley, M. (1989). *The dilemma of qualitative method : Herbert Blumer and the Chicago tradition*. London: Routledge.

Hansen, L. K., & Rieger, L. (2019). Interpretability in Intelligent Systems – A New Concept? In (pp. 41-49). Cham: Cham: Springer International Publishing.

Heffernan, T. (2020). The Dangers of Mystifying Artificial Intelligence and Robotics. *Toronto journal of theology, 36*(1), 93-95. doi:10.3138/tjt-2020-0029

Herschberger, C. (Writer). (2021). Artificial Gamer. In N. o. P. Milkhaus (Producer).

Hertfordshire, U. o. (2005). Kaspar. In: Adaptive Systems Research Group.

Hong, K., Chalup, S. K., & King, R. A. R. (2014). Affective Visual Perception Using Machine Pareidolia of Facial Expressions. *IEEE transactions on affective computing, 5*(4), 352-363. doi:10.1109/TAFFC.2014.2347960

Huber, P. J. (2002). [What Is a Statistical Model?]: Discussion. *The Annals of Statistics, 30*(5), 1289-1292. Retrieved from http://www.jstor.org/stable/1558711

IndigoGaming (Writer). (2020). Cyberpunk Documentary PART 2 | Ghost in the Shell, Shadowrun, Total Recall, Blade Runner Game. In.

ITU. (2018). Sophia. In. Flickr.com: ITU Pictures.

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2020). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI.

Kageki, N. (2012). An Uncanny Mind [Turning Point]. *IEEE robotics & automation magazine, 19*, 112-108.

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology, Vol. 6*, 360.

Katz, J. (2001). Analytic induction. In N. J. Smelser & B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (pp. 1--480).

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems, 134*, 113302. doi:10.1016/j.dss.2020.113302

Kim, M.-S. (2019). Robot as the "Mechanical Other": Transcending Karmic Dilemma. *AI & society, Vol. 34*, 321-330.

Knowles, B., & Richards, J. T. (2021). The Sanction of Authority: Promoting Public Trust in AI.

Kohs, G. (Writer). (2017). AlphaGo. In.

Kubrick, S. (Writer). (1968). 2001: A Space Odyssey [Live Action Film]. In. United States of America.

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Cent Rep, 49*(1), 15-21. doi:10.1002/hast.973

Marshall, F. (Writer). (2014). The Man vs. The Machine. In.

Marshall, M. N. (1996). Sampling for qualitative research. *Fam Pract, 13*(6), 522-526. doi:10.1093/fampra/13.6.522

MaxOfS2D. (2019). High-res 7.23 map render + GIF comparison with 7.20. In.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics, 30*(5), 1225-1310, 1286. Retrieved from https://doi.org/10.1214/aos/1035844977

McKee, A. (2003). *Textual analysis : a beginner's guide*.

Merzmensch, V. A. (2020). Pareidolia of AI. Retrieved from https://towardsdatascience.com/pareidolia-of-ai-dba7cf44bfde

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE robotics & automation magazine, Vol.19*, 98-100.

NSD. (2021). Fylle ut meldeskjema for personopplysninger. Retrieved from https://www.nsd.no/personverntjenester/fylle-ut-meldeskjema-for-personopplysninger/

O'Hara, K. (2020). Explainable AI and the philosophy and practice of explanation. *The computer law and security report, 39*. doi:10.1016/j.clsr.2020.105474

Oshii, M. (Writer). (1995). Ghost in the Shell [Animated Feature Film]. In. Japan.

Pereira, L. M., & Lopes, A. B. (2020). *Machine Ethics : From Machine Morals to the Machinery of Morality*(1st ed. 2020. ed., Vol. 53).

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137-141. doi:10.1007/s11747-019-00710-5

Ratcliff, D. E. (1994). Analytic Induction as a Qualitative Research Method of Analysis.

Rhee, J. (2013). Beyond the Uncanny Valley: Masahiro Mori and Philip K. Dick's Do Androids Dream of Electric Sheep? *Configurations, Vol. 21*, 301-329.

Ribeiro, J. (2016). AlphaGo's unusual moves prove its AI prowess, experts say. Retrieved from https://www.pcworld.com/article/420054/alphagos-unusual-moves-prove-its-ai-prowess-experts-say.html

Romportl, J. (2015). Naturalness of Artificial Intelligence. In (pp. 211-216). Cham: Cham: Springer International Publishing.

Romportl, J., Zackova, E., & Kelemen, J. (2015). *Beyond Artificial Intelligence : The Disappearing Human-Machine Divide*(1st ed. 2015. ed., Vol. 9).

Sanders, J. (2019). Half of employees think the cloud is actually in the sky, according to a third of IT workers. Retrieved from https://www.techrepublic.com/article/half-of-employees-think-the-cloud-is-actually-in-the-sky-according-to-a-third-of-it-workers/

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks, 2*(6), 459-473. doi:10.1016/0893-6080(89)90044-0

Sasaki, K., Ihaya, K., & Yamada, Y. (2017). Avoidance of Novelty Contributes to the Uncanny Valley. *Frontiers in Psychology, Vol. 8*, 1792.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International journal of human-computer studies, 146*. doi:10.1016/j.ijhcs.2020.102551

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489. doi:10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354-359. doi:10.1038/nature24270

Stein, J.-P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition, Vol. 160*, 43-50.

Strait, M. K., Floerke, V. A., Ju, W., Maddox, K., Remedios, J. D., Jung, M. F., & Urry, H. L. (2017). Understanding the Uncanny: Both Atypical Features and Category Ambiguity Provoke Aversion toward Humanlike Robots. *Front Psychol, 8*, 1366-1366. doi:10.3389/fpsyg.2017.01366

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning : An Introduction*.

Švarný, P. A Visit on the Uncanny Hill. In (pp. 133-142). Cham: Cham: Springer International Publishing.

Taulli, T. (2019). *Artificial Intelligence Basics : A Non-Technical Introduction*(1st ed. 2019. ed.).

Troshani, I., Rao Hill, S., Sherman, C., & Arthur, D. (2020). Do We Trust in AI? Role of Anthropomorphism and Intelligence. *The Journal of computer information systems*, 1-11. doi:10.1080/08874417.2020.1788473

Tu, Y.-C., Chien, S.-E., & Yeh, S.-L. (2020). Age-Related Differences in the Uncanny Valley Effect. *Gerontology, Vol. 66*, 382-392.

Turing, A. M. (1950). Computing machinery and intelligence. In (pp. 40-66). Oxford: Oxford University Press.

Van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology, 54*, 93-100.

Weller, A. (2019). Transparency: Motivations and Challenges. In (pp. 23-40). Cham: Cham: Springer International Publishing.

Windsor, M. (2019). What is the Uncanny. *The British Journal of Aesthetics, Vol. 59*, 51-65.

Zackova, E. (2015). Intelligence Explosion Quest for Humankind. In (pp. 31-43). Cham: Cham: Springer International Publishing.

Znaniecki, F. (1934). *The method of sociology*. New York: Farrar & Rinehart.