UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Games and AI: How did game engines become unbeatable and does AI dominance entail any risks

Mohammad Odeh

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

dr. Sebastian ROJAS GONZALEZ

UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2022
2023

UHASSELT

KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Games and AI: How did game engines become unbeatable and does AI dominance entail any risks

Mohammad Odeh

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

dr. Sebastian ROJAS GONZALEZ

Acknowledgments:

I would like to thank my supervisor Mr. Sebastian Gonzalez for giving me the right combination of guidance and autonomy to investigate the research questions that have been of interest to me for years before starting this master's program. Many thanks go to my family for their full support, to my father for encouraging me to enroll in a master's program, and to my mother for instilling in me a tireless sense of determination when I set my sights on any goal. My thanks also go to "The Oddz" for making this ride easier, to the good friends I made in Belgium, and to whoever takes the time to read this thesis. Special thanks to Hind Kh. for helping me design the defense poster.

-Mohammad Odeh

Abstract:

The rapid advancements in Artificial Intelligence applications could prove to be the best event for humanity or would pose an extinction risk or fates worse than that according to some, such as prominent scientists like Stephen Hawking and AI experts like Geoffrey Hinton. Other intellectuals warned of unchecked technological advancements prior to the AI blooms. However, some of the extreme possible risks are not rooted in actual events as much as in speculative fictional scenarios or misconceptions about computer vs human games. To separate the realistic wheat from the superstitious chaff a proper dive into AI history and capabilities is a must. This thesis seeks two main objectives, the first is to peer at the history of AI in general and of games in particular, as games that require strategizing were considered to be a hallmark of human intelligence. The second is to carry those conclusions as premises and review further literature to better understand and categorize the risks posed by AI. The thesis proposes the idea that AI advances in 3 stages: Tool - Adversary - Guide (T.A.G) and that the Adversary stage is part of human design not inherent in AI. The thesis attempts to relate risks to each other instead of reviewing them separately and proposes a model to link the risks together. Both contributions are meant to be challenged and enhanced with further research to become robust enough and rationally inform R&D practices and policy decisions.

Contents

Acknowledgments: 1
Abstract: 2
Introduction**:** 4
Methodology, Research Questions & Limitations: 5
Definition: 7
Chapter 1: Brief History of Artificial Intelligence: 8
Game AI history 11
Chapter 2: AI rise to game championships: 13
Dice: Backgammon 13
Draw: Checkers 14
Defeat: Chess 17
Double Tap: Go 22
General remarks: 24
T.A.G Stages: 24
Champions POV: 25
Guides to Gods: Apotheosized Intelligence 26
Chapter 3: AI is king, now what? The potential risks of using AI. 27
Risk I: The perils of prediction in AI: 28
Summary: 31
Risk II: Anti-Tech Views & Actions, Questions Regarding Technology 32
Summary: 37
Risk III: Current and Concrete Risks 38
1. Economic setbacks 38
2. The Alignment Problem: 40
3. Military AI: 41
4. Epistemic Regression 43
Overlapping Risks: 45
Summary: 46
Risk IV: Speculative Future Risks, welcoming our AGI overlords. 47
The risk of worrying about risks: 47
The pincer of regulation and competition: 49
Summary: 49
Conclusion: 51
References: 53
Appendices: 63
Appendix 1: T.A.G Stages Summary. 63

Introduction:

Advances in Artificial Intelligence (AI) and technology overall is reshaping our lives at a dizzying pace, in the 90s the word "internet" had to be explained on TV interviews, today we expect Wi-Fi waves to fill our surroundings like oxygen.

The relationship between humanity and its creation is becoming as complex as the creation itself, it seems logical that all tools created by humans were created for the benefits yielded when such inventions are wielded, even the most destructive tools such as atomic bombs give their owners an advantage over their foes. While the accidents of such advancements could not be understated, examples include the nuclear disasters in Chernobyl in 1968 or Fukushima more recently in 2011, the threats posed by AI are of a different magnitude as they stem from the ultimate risk that a sentient super intelligence may act against the human creators, plus a myriad of risks that are more realistic.

While many previous tools and technologies surpassed human capabilities, like a crane lifting tons or an airplane defying gravity, these inventions were never worrisome in the sense that they will never willfully act against the interest of their operators, and any possible damage is accidental not deliberate by the tool itself. Yet AI poses risks even when the tool is functioning according to the orders as it understands them when that understanding is not aligned with the users' intentions.

The promises and threats that could accompany seemingly sentient creations existed in ancient poems and recent science fiction media, lately as some promises are realized and with threats looming, AI is scaring many experts of the field. Even AI proponents and tech experts call for proper direction, in an open letter that claims AI has been steadily improving over the last few decades and it promises grand economic and social benefits for tweaking different AI designs, it cites many fields as examples; speech recognition, image classification, autonomous vehicles, and question-answering systems. The letter calls for research to make AI robust and more beneficial so it can head in the right direction. It was signed by thousands and notably by forerunners in AI research and tech CEOs. (Future of Life Institute, 2015).

After 7 years many big milestones and specifically the advancement of large language model AI chatbots such as ChatGPT transformed the call for safely directing AI research to a call for slowing down research to a complete halt, as another open letter from Future of Life Institute (2023) did, arguing that we should consider the threat of AI outsmarting us, or being utilized to spread propaganda. The letter ultimately wants AI development to make sure that development is contingent on a net positive; that the effects are positive and the risks manageable.

The risks are serious enough that they led Geoffrey Hinton, nicknamed the "the godfather of AI", to resign from Google so he could speak freely about AI, much like the previous open letter, his main concerns seem to be the propaganda potential of AI and the risk of AI outsmarting humans. (Taylor & Hern, 2023).

These statements taken alone, may give the impression that these are critics of AI but it's important to observe that none of them have gone to a radical conclusion of banning AI altogether, a call for a pause is as extreme as it gets. This doesn't mean that there are no thinkers who ask more radical questions about the way technology advances, it may be easy to view them as neo luddites that want to destroy technology, but that view doesn't give this camp the credit that its due. Some pf the intellectuals already warned against unchecked technological progress, thinkers like Jacques Ellul in his books "Technological Society" and "Technological Bluff" that question technological autonomy and impacts of technology, Lewis Mumford in "Myth of the Machine" that introduces many ideas such as the Mega Machine and the Pentagon of Power, and Martin Heidegger in "Question Concerning Technology" which views technology as a method that turns nature into a standing reserve.

A thorough discussion of such ideas is outside the scope of the thesis but a brief presentation of some of their ideas will help explain the risks that AI poses. They all had something to say about the way technology warps our reality, senses, and nature around us. Just because the larger anti-technology

camp has no Musk type billionaires shouldn't give us the illusion that there are no valid points against technological advancements without falling into primitivism.

The last chapter of this thesis will critically assess and categorize the risks and the differing points of view, but before getting into the mainstream scares of a Skynet future the thesis will start with a chapter on the history of AI to outline the context, as the ever-changing nature of this technology is too complicated, and the updates are getting more and more rapid. Lessons must be learned from its rich but brief history. This thesis will limit its scope to the field of games for the following reasons.

The main scare of AI outsmarting us is tied with science fiction, but it only became more realistic when this tool exhibited an ability to surpass humans, the most famous example being the 1997 match between chess grand master Garry Kasparov and IBM Deep Blue computer. 3 years prior, the checkers grand master emeritus, Marion Tinsley, had a streak of draws against Chinook. Games that are less complicated had similar encounters much earlier, in 1979 the world champion in backgammon, Luigi Villa, lost to BKG 9.8 which was the first time a human champion loses to a machine. It was a matter of time before complex games like Go had the fated achievement when AlphaGo program defeated the number one player of Go, Lee Sedol in 2016.

The wins for such computer players shifted the way players view these tools, they are no longer unheard-of programs but are now used by players to learn the game and improve their gameplay. They have long started solving problems in their respective games (Tesauro, 1995; Schaeffer, 1997; Hsu, 2002). Thus, we can see that the wins did not abolish the games between humans, this thesis will draw a preliminary conclusion of how the dynamics between humans and game AI progressed from mere tools to adversaries to guides (abbreviated as T.A.G) then this cycle will be further questioned and refined to address the possible AI risks.

The guide stage synergy is exemplified by Centaur Chess which allows teams of human players and chess computers to work together, or as Bridle (2018) points out, the average human player with an average computer may beat top human players or supercomputers. Bridle concludes with this remark, while there are many risks involved with AI, we can't dispense with this technology, it is up to us to steer its progress, to reach what I call the Guide stage.

The guiding and the help in decision making is not limited to games, humans rely daily on search engines and social networks that apply AI to get better results, the web cookies and online advertising services are another example, and so are autopilots, traffic control software, automated phone answering services. Such applications fall under the category of Artificial Narrow Intelligence, while more General Intelligence and Super Intelligence are not yet achieved (Gurkaynak et al., 2016) but will be addressed in the thesis.

In short, this thesis aims to synthesize literature and attempt to draw a full picture of the following points in the AI journey. 1st chapter will present a general history of AI, the 2nd chapter will focus on game AI and track the progress towards the threshold of AI beating humans in board games like Chess, Checkers and Go, and what those breakthroughs tell us about the nature of AI and human competitiveness. This should give an idea of other more natural purposes of AI as a tool and how humans start depending on it, and how it fulfils its primary function of informing our decisions rather than acting as an adversary. And the final chapter will propose a model after a critical review of the claims from AI enthusiasts and critics, AI will be examined both as a special technology and as the latest extension of general technological and industrial advancements.

Methodology, Research Questions & Limitations:

Methodology

The thesis is of the literature review type. Literature review is an important step in all kinds of research as it helps situate the problem, to help others avoid reinventing the wheel and to build models and hypotheses to be tested. As research mounts it is important to take a step back and

review existing literature to synthesize, to facilitate future research and to come up with further questions that could be investigated in the future.

The method followed in this thesis was a selection and study of 5 books (Schaeffer, 1997; Hsu, 2002; McCorduck 2004; Ford, 2015; Christian, 2021) as main sources for the 3 chapters, then going into the rabbit hole of citations to follow traces of a general arguments regarding Human and AI relationship with a focus on games and risks, to critically review each source and weave a coherent thread of all the presented facts and ideas. Peer reviewed papers in respectable journals were the main source. Due to the rapid pace of AI in recent times the citations also include many news articles, few think pieces and blogposts to present different kinds of arguments and concepts.

The books were selected based on 3 criteria: relevancy (to the topic), familiarity (to the thesis author) and standing (of the book). To take the game chapter and an example of relevancy, one of the main events that come to mind is the famous Kasparov-Deep Blue 1997 match, so it seemed logical to start with reading a book by one of the main programmers of Deep Blue, Feng Hsiung Hsu, and to do the same in the less famous case of checkers, and so two sources that check relevancy and standing were selected. When it comes to standing then Pamela McCorduck's book on history of AI is frequently cited and the author of this thesis was familiar with it prior to working on this paper.

As will be shown in the 3rd chapter, AI poses epistemic and educational risks, which might warrant the following disclaimer: The author of this thesis did not use any Large Language Models (LLM).

Research Question

This thesis seeks 2 main objectives, first to present the indivisible interrelation between AI and games and then to build on it to hypothesize a pattern of human AI dynamics, and second is to review the risks that are posed by reliance on AI. To formulate the objectives in the form of research questions they would be:

1. How did AI advance in board games?
   1a. What does that tell us about human and AI dynamics?
2. What risks does AI dominance entail? (Depending mainly on the answer to the 1st question)

It might be argued that the objectives could be done separately in different papers but to understand the risks of AI it is important to understand its history and the history of AI is not separable from its application in games, as the following chapters will show, programmers aimed to use computers to play chess since the beginning, way before AI emerged as a field in the 1950s. And more importantly the common conception of AI risks has been shaped by our understanding of how AI successfully dominated many games, I will argue that this gives a wrong impression about how adversarial AI really is, without discounting the real risks.

As for the importance of this thesis objective and the risk chapter in particular, at the time of writing this thesis there were rapid developments of multiple AI applications that threatened many jobs, art generating AI threatens the creative field which was previously marked safe. LLMs are so powerful that prominent researchers are ringing the alarm bill about AI capabilities growing without the proper guardrails, and open letters are asking for a halt in AI research. If many people of worth believe that AI may threaten humanity at large, then the importance of critically reviewing the uses and history of this tool couldn't be understated.

The intended contribution of this thesis is to firstly present the game AI history and secondly to give a somber categorization of the risks. Finally, if my conclusions are wrong, the thesis provides the raw facts in case the reader wants to draw different conclusions.

Limitations

The thesis is limited for multiple reasons. The author is not a computer scientist, and his background in Mechanical Engineering is not directly related to the topic, the insights of a programmer would have enriched this thesis from a technical perspective. However, the author is interested in the relation between humans and technology from a social and ethical point of view prior to picking the topic and this interest drew him to it, this helped because the literature review did not start from scratch.

The author also happens to be a fiction writer which may lead to this thesis sounding more like chapters in a book than a master's thesis, this could be a bug or a feature depending on the reader's preference.

The pace of AI development may render some of the findings outdated quicker than expected, one may hope that some solutions to some of the risks in the 3rd chapter will be available soon, and one fears that more novel risks will appear sooner than expected.

The limited scope of gaming AI may not present the full picture, similar research should be done in all fields of AI with an eye on what it tells about the development of this tool, the dynamics it has with humans and what may be inferred from that to help check the possible and real risks posed by AI.

Definition:

The term "AI" will appear over a hundred times in this thesis, many definitions of the term Artificial Intelligence exist, for the purposes of this thesis the main definition will be that of John McCarthy who coined the term back in the 50s, in a more recent Q&A with him he defines AI as "the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable." Intelligence in this context is defined as "the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals and some machines." (McCarthy, 2007, p.2)

In some places the word will refer to the applications of that science in the form of machines or programs rather than the science as a field, the human intelligence of the reader is sufficient to tell them apart contextually.

Some believe that there are future levels that AI will reach, Gürkaynak et al. (2016) start their paper with definitions of all 3 levels, the next level would be Artificial General Intelligence "AGI" which refers to a level at which AI becomes as good as human intelligence, and the 3rd level would be Artificial Super Intelligence "ASI" which exceeds human intelligence and might result from AGI improving itself in what is known as the singularity. However, AI have only exhibited the first level so far, at the present state it could be called Artificial Narrow Intelligence "ANI" since it has a narrow field or task that it's specialized in, thus "AI" will be referring to "ANI" unless otherwise stated.

Kaplan and Haenlein (2019) also start their paper with a good overview of previous definitions of AI, but they attempt to craft their own definition based on 3 different types of human intelligence, but as the 2nd chapter will show playing to the computer computational strengths proved superior to mimicking human reasoning, which is why their definition may work better for speculations about future types, but is inadequate when it comes to understanding the history of machine and artificial intelligence.

Chapter 1: Brief History of Artificial Intelligence:

This chapter serves to draw a rough timeline of the evolution of AI to situate the scope of the thesis within, and to give the two main chapters a larger context. First it will represent McCorduck (2004) timeline supplemented with further information to give the reader an overview, then focus on AI and games timeline according to Cirasella and Kopec (2006) to provide context for the game chapter.

According to Pamela McCorduck, the history of AI is entangled with the history of data collection, programming, computers, and human conception of manmade creations that think autonomously, in her book "Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence" which is the primary research material for this section, she traces the history of collecting information into libraries and encyclopedias as far back as ancient Babylonians and Assyrians centuries BC.

Then she moves to the notion of automata which is found as early as in the Homeric poems, in book 18 of the Illiad where Hephaestus had 20 bellows that work on verbal command and robot-like creatures that have understanding and speech abilities, and in book 8 of the Odyssey there are fast auto piloted ships provided by the Phaeacians, the ships can navigate without maps, requiring only the input of a destination.

When not done by thinking ships, sea navigation could suffer from costly and even fatal human errors in manually inked tables, according to McCorduck (2004) Charles Babbage built a small model of the "Difference Engine" in 1822, an automatic calculator that can finish tables essential to navigation and ballistics, he persuaded the British government to finance a larger model, however the project was too ambitious, hindered by primitive machining state, and Babbage being distracted by other plans.

Although it was never realized, it is estimated that the effects the progress of creating the engine had on machining covers the huge investment by the British government that withdrew the support.

Later Babbage conceived of the Analytical Engine, a grand all-purpose calculator capable of analysis and tabulation of any function, together with Ada Lovelace according to some historians, they worked on the new engine, but it was too costly, and it too was never completed. In the context of this thesis, it is worth noting that they considered having it play tic-tac-toe and chess and developed a system for horse betting. (McCorduck, 2004) This shows how interconnected computers and AI with games are.

In 1944 ENIAC (Electronic Numerator, Integrator and Computer) comes online at the University of Pennsylvania, in 1949 Mark I, the first stored-program computer, comes online at Manchester University, Alan Turing attempts to program it to play chess, in fact one of the very first papers about AI is by none other than Alan Turing (1950), called by some the father of AI and computer science, the paper attempts to answer the question "Can Machines Think?" and proposes playing chess as a benchmark, the importance of this paper merits more exploration here.

The paper attempts to answer by setting a thought experiment called the imitation game, the game has 3 participants, one interrogator has to discern the gender of two players that are not visible to him and will answer his questions only through letters penned on a typewriter, then the game is updated by Turing to be played with a machine in the place of person answering, the goal becomes to tell if the participant is a man or machine, with the machine attempting to pass as man and the man playing normally. Again, we can see how games and AI went hand in hand right from the start.

Turing continues to fine tune the experiment to specify a discrete-state digital computer for machine and reformulates the question to make it: "Are there imaginable digital computers which would do well in the imitation game?" (Turing, 1950, p. 442). Afterwards he addresses 8 objections to the idea that machines can think; *theological* regarding the souls, *avoidance* of the problem for its dreadful conclusions, *mathematical* limit of the machine answers, *unconsciousness* or the lack of depth in artistic depth in machines, *disability* of the machine in one form or another, *unoriginality*, *discreteness*

which makes it not as continuous as the nervous system, *formality* of machines that makes it hard to mimic human ability to respond to novel situations, *extrasensory* phenomena that may interfere with the proposed game.

He doesn't consider all with the same seriousness but attempts to offer counter arguments to each, and concludes the paper with the idea of a child like computer, that is a computer that can be programmed like a child and it can learn gradually to become a computer that can think like us. Finally he entertains the game of chess as an example of an abstract activity that's good for a starting point to test machines ability not only to think but to compete with humans.

The abstraction in chess is one of the many reasons that made it and similar games interconnected with AI research since the beginning, the history of AI and the dynamics between humans and AI cannot be fully understood without a look into the game aspect.

The Turing Test is a prominent example of measures for reaching artificial intelligence, however I believe that Goodhart's law may apply to it, the law states that when a measure becomes a target is when it stops to be a good measure, in other words a machine may pass the test without exhibiting any other features of intelligence, it will be another case of narrow AI which dominates in one field but is useless the minute it crosses the boundaries of that test to other tests, or it may be that this test doesn't assess intelligence and AI might become objectively intelligent regardless of its ability to play this particular game.

One of the major counter arguments to this line of thought was presented by John Searle in his book Minds, Brains and Science (1984), in general his counter argument to the AI community views that a program could think like humans because it can manipulate symbols is that human thoughts have semantics not just syntax without a meaning, he presents his counter thought experiment known as the Chinese room. It must be pointed out that Searle didn't present it to directly counter the Turing test but to counter the premise that an outward human like behavior proves an existing internal human intelligence.

The thought experiment asks us to picture ourselves in a room where we receive cards with incomprehensible symbols on them, and from a basket of cards we are meant to match it to another specific card then send that one to another room. If those symbols were Chinese letters constructing a question, and if the matching process that was given to us was good enough to yield answers to those questions just like a native Chinese speaker, then from the outside it may be assumed that the person inside is a native speaker, but he does not understand the symbols he is shuffling.

This argument in turn also has many counter arguments that vary in their strength, "The Systems Reply" which states that even though the man in the room doesn't understand the language, the system as a whole does. "The Robot Reply" that argues that a robot with sensors and mobility would get the semantics. Other replies include the Brain Simulator, Other Minds and the Intuition Reply, although it is an interesting topic with larger philosophical implications, this chapter is focused on the AI history. Cole, D. (2004) does a great job at detailing the Chinese Room argument and the replies mentioned above.

The paper by Turing is a foundational for its test proposal, it spurred many papers to discuss the interpretation of the game, Piccinini (2000) lays out the standard understanding of the test, which is the one mentioned in this thesis, and argues that it's a better fit than a literal understanding, that the machine is playing in lieu of the male player. On the opposite side Traiger (2003) finds flaws in different interpretations of the test and argues that what Piccinini called the literal understanding is stronger than the standard test. While it is not accepted by everyone to be the true benchmark of intelligence or even a test that has a final say, its importance cannot be understated.

Hair splitting aside, the next marker in the AI timeline is the year of 1956 where two significant contributions were made in Dartmouth conference, the term Artificial Intelligence itself was proposed then and started propagating, and the program "Logic Theorist" made by three scientists, Allen Newell, Herbert Simon and J. C. Shaw was presented in the conference, it succeeded in proving

theorems in Principia Mathematica, and it was, according to McCorduck (2004), the first intelligent computer program.

The 50s also saw the production of mainframe computers, bulky large computers used by organizations for data processing, statistics, ERP systems and transaction processing, then in the 60s multiple programs successfully expressed levels of intelligence such as STUDENT which understood natural language enough to solve algebra word problems, SHRDLU was an early language parser that could communicate with the user about a virtual world with blocks that it can move around and answer questions about. DENDRAL was the first knowledge-based program for scientific reasoning and MACSYMA was the first knowledge-based program in mathematics, and ELIZA which was a natural language processing to simulate a conversation with a human, it acted as a psychiatrist and was convincing for its time.

In 1977 a Stanford cart that was originally meant to simulate a moon rover was rebuilt by Hans Moravec, who added a camera that took pictures from different angles to send them to a computer that calculated the distance, two years later it managed to autonomously cross a room filled with chairs.

The 80s encompassed the development of personal computers, another development in 1983 was a book titled "The Policeman's Beard is Half Constructed", published by Mindscape, it was written not by a human author but by an AI program called "Racter" that was in turn written by William Chamberlain and Thomas Etter, which makes it the first novel of this kind. The following year a commercial version of the program was released, Dewdney (1984) described Racter as a schizophrenic program that produces grammatically sound, funny, and somewhat nonsensical text. Dewdney compares it to ELIZA and SHRDLU, and he considers SHRDLU to be the most convincing AI interlocutor of the three, because it limited its scope to object arrangement whereas the other two attempted to comprehend many kinds of dialogues.

During the 70s and 80s there were serious cuts in funding for AI projects and was subsequently called "AI winter", Haenlein & Kaplan (2019) compare the history of AI to the 4 seasons, with the 40s being the beginning of the spring, the summer following it until the winter that was ushered in by many factors, the main one being the failure of AI to give results that were promised in spring time. Garvey (2018) presents more details about this AI cycle.

Two words can explain how AI winters start: "Broken Promises". The 3rd chapter will expand on this idea but here the following example would suffice. The British Science Research Council commissioned the British mathematician James Lighthill to write a report that quizzed the optimism of AI researchers. The year was 1973 and back then he concluded that machines would only reach the level of experienced amateurs in games like chess, and that AI will never be able to do common-sense reasoning. This led to cuts of support for AI research by British government in most universities, then the U.S. government followed which started an "AI Winter". An interesting fact worth mentioning is that Lighthill James occupied Cambridge University's Lucasian Chair of Applied Mathematics, the holders of Lucasian Chair are asked for advice by the government, and Sir George Biddell Airy who was such a holder, advised Queen Victoria against continuing support for Charles Babbage's difference engine. (Crevier, 1993)

Crevier (1993) in his history of AI writes about expert systems in AI winter, they were growing larger with steady increases in rules without necessarily becoming more efficient, such systems could not acquire knowledge by themselves but required spoon-feeding information, an example of the diminishing returns of these systems can be gleaned in 1987 when Teknowledge and Intellicorp, two of the leading expert system shell developers, lost what amounts to more than $6 million.

More examples of setbacks provided by Crevier (1993) in AI winter included the failure of certain projects to meet their goals, in mid 70s the US Defense Advanced Research Projects Agency (DARPA) cut funding for project "SUR" short for Speech Understanding Research, the project cost 15 million dollars within five years, and the results didn't satisfy DARPA despite the claims of the former project manager that the benchmarks were met.

A decade later the U.S. military financed a project of a "Smart Truck," an Autonomous Land Vehicle project starting in 1983 as part of ten-year Strategic Computing Initiative program, The truck was designed for missions such as weapons delivery, reconnaissance, ammunition handling, and the project accounted for about a quarter of the program's $100 million annual budget. Hans Moravec argued that it was too early for such a feat, and in 4 years the failure of the project became clear.

Things started to look up in the 90s, the main source for the following milestones moving forward is Frana and Klein (2021), game supercomputers managed to draw and defeat world champions in games such as Chess, Checkers and Othello. Moreover, there were advancements in robotics, the U.S military started using unmanned aerial vehicles such as General Atomics MQ-1 Predator, Sonny launched a robotic dog called AIBO and commercial speech recognition software were available such as NaturallySpeaking.

During the 2000s DARPA was still seeking autonomous vehicles, in 2004 it launched challenges like the Grand Challenge with a 150-mile course but none of the contestants managed to finish it, in 2007 it launched the Urban Challenge that tests autonomous vehicles' ability to handle traffic. On the commercial side, Netflix offered a million-dollar prize for developing a recommendation system that is 10% better than their own system Cinematch, it was a tight competition between two teams, but the winner was the multinational team "BellKor's Pragmatic Chaos" (Lohr, 2017)

In the following decade there were state of the art advancements in many fields, some successful as Siri, which exemplifies voice recognition in virtual assistants, others like Tay, a chatbot that succeeded from a technical perspective in learning from its interactions with human users, but it raised an alarm about the bias introduced by such interaction, it was taken down after imbibing enough taboos.

In 2016 the game of Go was dominated by AlphaGo and then more advanced game engines such as Alpha Zero were developed and trumped multiple game programs without having to learn from human experience. Even Poker didn't escape the AI onslaught when Libratus won against top human players.

From 2020 and up to the time of writing the thesis there were even more fast paced advances and accordingly louder voices against the neck breaking speed of AI development, two new fields that seemed to be safe in the past are seeing the AI foot in their door, the first being art generating AI software that creates high quality images from text prompts using stable diffusion. The output quickly became good enough to pass a visual Turing test, most viewers of AI generated imagery cannot tell if it's real or not. In the case of artworks it's becoming hard to tell if they are created by humans.

In 2021 DALL-E was released by OpenAI and so was a newer version DALL-E 2 a year later, initially it produced bizarre and comical images but it rapidly improved and this was the case for similar AI applications such as WOMBO dream and MidJourney. The voices were also subject to manipulation, in 2022 ElevenLabs released text-to-speech AI software, that can mimic natural speech including the emotion and intonation, since it learns from data it can replicate specific individuals voices which opened the flood gates to parodies ranging from the comical to the horrifying, the combination of these tools raise copyright concerns and an epistemological question mark that will be discussed in the 3rd chapter. One such concern is AI crafted songs that use voices of famous artists (Mendez, 2023).

OpenAI also released ChatGPT in March of 2023, an AI chatbot using a LLM and is able to produce a wide variety of texts from essays to business pitches to lyrics and poetry, it can also create basic programming codes, and while it's not perfect as it sometimes hallucinates fake citations and may be jailbroken to give harmful answers, it is a huge step for AI, and it's too early to grasp the extent of the possible costs and benefits we gain from such software.

Game AI history

As stated earlier this brief history of AI from antiquity until 2023 serves to contextualize the discussion of human AI dynamics in games and the emerging risks of reliance on AI, before moving to the main 2 chapters of the thesis, a compact overview of AI in game history will be provided using Cirasella, J., & Kopec, D. (2006) as a main source up to 2006.

The relationship between AI and games is inseparable as games were some of the initial projects for computer scientists worked on to develop the field, between the 50s and the 60s Alan Turing and Claude Shannon wrote the first algorithms for chess programs, then Alex Bernstein wrote the first fully functional chess program, whereas Tom Throop wrote the first program for playing Bridge a few years after Arthur Samuel wrote the first checkers program.

The 70s started with Al Zobrist writing the first Go program, Nicolas Findler writing the first Poker five-card draw program, as for chess the slow but steady advances started. CHESS 4.5 broke the USCF expert rating, and another checkers program was written by a team from Duke university and it was able to beat Samuel's program but it was no match for human grandmasters, right before the 80s the first milestone in man vs machine games was passed when Hans Berliner's program BKG 9.8 had a humble win against Luigi Villa in an exhibition match.

In 1980 Mike Reeve and David Levy wrote the Moor program for playing Othello, it scored one win out of 6 matches with the world champion Hiroshi Inoue, Go programs were too weak to play against humans but that didn't stop computer Go tournaments from being held, in another computer tournament, the international Computer Olympiad, a Backgammon program written by Gerald Tesauro won. Of course, chess programs were improving as well. Ken Thompson's Belle broke the chess master rating of 2000 in 1983 and five years later Hans Berliner's HiTech broke the 2400 rating, but computers were still not good enough to defeat the champion Garry Kasparov who won against Deep Blue's predecessor Deep Thought in 1989, one year after Deep Thought was the first computer to defeat a human grandmaster in tournament settings.

The 90s witnessed multiple milestones being reached by game programs and computers, the most famous is the 1997 match between Kasparov and Deep Blue, but it wasn't the only program that dethroned the human champion, in checkers the title was won by Chinook and it held the title until it was retired in 1997, that year also held the match where Logistello, a program by Michael Buro that generated its expertise by playing against itself, won against the world champion Takeshi Murakami.

Less flashy events were also taking place, in the early 90s Zia Mahmood, a champion in bridge, offered a one million dollar bet to any program that can beat him and in 1998 a program written by Matthew Ginsberg called GIB finished 12th in World Bridge Championship, and it played against Zia Mahmood in an exhibition match, with a performance good enough that he withdrew his bet.

After surviving Y2K there were more milestones to be achieved by game computer programs, multiple tournaments that were held like the 2004 Man vs. Machine World Team Championship that resulted in 8.5-3.5 for the computers, then between 2010s and Corona some of the most difficult games such as Go were also dominated by AI, in 2011 IBM's program Watson won Jeopardy! against two champions (IBM Corporation, 2011), in 2015 and 2016 Alpha Go won first against Fan Hui then against Lee Sedol, in 2017 Libratus, a poker playing program managed to beat 4 human masters for about $1.77 million dollars. (Ghose, 2017)

There are other ways to view the history of AI, as Simos, M et al. (2022) argue that AI timeline could be split in two pieces distinct in terms of materialities and discourses, the first starts with the introduction of computers and the second with the ubiquitous use of computers and the internet. However, the article in my opinion does not present enough evidence to back up such a clear-cut distinction between the two periods, but it does present a good analogy of AI as steam power.

Chapter 2: AI rise to game championships:

The reason board games were appealing for the early AI programmers was that they had suitable conditions to be investigated, they were summed properly in an early paper by Samuel (1959) on why he chose Checkers, the 5 conditions were that games are deterministic yet not solved already, there is a definite goal or reward, the rules must be definite, there exists enough knowledge about the game for testing and that enough people are familiar with the game.

This chapter of the thesis will provide an overview of AI in games with a focus on the biggest achievement, that is defeating a human champion. The games covered in detail are Backgammon, Checkers, Chess and Go in chronological order of relative success against Masters of each game. These games were chosen because each presents a sample of the AI progress in different methods and outcomes, each part will summarize the relevant key literature and provide a brief view of the evolution of programs that won the final challenge or had an impact on the progress towards it.

This chapter attempts to back up the claim that through judging the field of games we find 3 stages in Human-AI dynamics; from tools to adversaries to guides (T.A.G), the chapter showcases patterns that emerge from the literature review and the common perceptions of AI versus the historical reality.

Dice: Backgammon

Berliner (1980) documents the first successful encounter for computers against human experts in a game that requires intelligence, on July 1975 Luigi Villa the backgammon world champion lost to BKG 9.8, a computer program developed by Hans Berliner who was also leading the HiTech chess computer team in Carnegie Mellon University, the program won 7 matches to 1 with $5000 were at stake.

Berliner admits that the analysis shows Villa to be the better player, luck did play a role in mitigating two mistakes made by the program, and postmortem analysis shows that Villa made decisions that are mostly better than what the program would have chosen if the roles were reversed.

Unlike other games that will be discussed the game of Backgammon has a luck element to it, each move depends on a throw of the dice but selecting the right move requires intelligence and knowledge, as the game has up to 10^{20} positions, Berliner argues that this makes it as complex as Checkers, if we factor in that the dice roll may result in 21 possible combinations, and the computers at that time couldn't search to a factor of 400 for each ply.

Tesauro (1995) also states that brute force search isn't feasible in Backgammon like it is in Chess or Checkers due to 21 dice combinations and 20 moves for each combination which branches to a higher number of possibilities than other games.

The following paragraphs summarize Berliner (1980) paper on the methods used, which considered features of the game in a feature hyper-space, such features could be as simple as the location of a piece or more complex, Berliner uses the blockage as an example of a complex feature, for successful blocking through the lens of this feature-based program the difficulty of breaking through the blockage should be evaluated.

The team programming BKG used what Berliner dubs SNAC method, an acronym for Smoothness, Non-linearity, and Application Coefficients. The term Smoothness refers to the curve of game states, in chess for example there is an opening phase, a middle game and an endgame phase. These states require different considerations, if the program isn't aware of the current phase, then it may rush some moves. This was solved by defining the endgame as a smooth function without sharp jumps. As for Non-linearity, it allows for more sensitivity and variation, finally, the Application Coefficients refer to variables that are constrained to get sensitivity without volatility.

This SNAC method made BKG 9.8 superior to its predecessors and able to compete with humans, the limitations of the match with Villa are firstly the fact that it is a limited exhibition match, it was not a tournament match as with the rest of the programs in this chapter. Villa may have not taken the program seriously due to its appearance and the weakness of commercial backgammon programs at the time.

The tale of Backgammon doesn't stop there, as Gerald Tesauro (1989) wrote a program called Neurogammon which uses multilayer neural networks that learned to play through backpropagation training on expert data sets, the networks have a hidden layer and a standard feed forward architecture, trained from a set of 400 games. It won the Computer Olympiad of 1989 in London, against 5 computer opponents: Video Gammon, Mephisto Backgammon, Saitek Backgammon, Backbrain, AI Backgammon. It also played against a human expert, Ossi Weiner, but lost the game after a praiseworthy performance according to its human opponent.

Another step by Tesauro (1995) was to forgo reliance on human experts because they are not infallible and their experience is subject to revision with time, so the program included networks in a standard Multilayer Perception Architecture (MLP) with a feed forward flow from input nodes to output nodes through hidden nodes, through nonlinear sigmoidal operation and backpropagation algorithm.

Initially the program was playing against itself without any special knowledge of the game, it was observed that the program first extracted linear evaluation function, it learns basic strategies and tactics without any prior knowledge.

With 40 hidden layers and 200 thousand games it was almost as good as Neurogammon. Then Tesauro added a set of features that define better gameplay, with that TD-Gammon surpassed Neurogammon. (Tesauro, 1995) this is as an early example of game AI as is known now, it's different from Neurogammon, DeepBlue and Chinook with its considerable self-training.

Although TD-Gammon didn't play an official championship match it did play 38 exhibition games against professional human players including world champion Joe Sylvester. Its final form TD-Gammon 2.1 that trained on 1.5 million games had a tight match with Bill Robertie, who is twice a winner in World Backgammon Championship.

Another top player Kit Wooslie ran extensive analysis and concluded that this program has an advantage over humans in positional play, he argues that its strength is in judgment of the position unlike chess programs that are good at tactical calculation but not at positional play. Tesauro (1995)

While the previous information could be used as an example of the adversary stage in the T.A.G trio, the next stage of AI as a guide is exemplified by a change in gameplay in one of the openings, one particular opening roll called "slotting" was the favored choice of experts to "splitting", however TD-Gammon made "splitting" more common among human experts after analyzing this opening.

TD-Gammon was a step towards self-training programs that culminated in AlphaGo almost two decades later and BKG 9.8 was the harbinger of things to come in terms of defeating human champions, the next game computer, Chinook, couldn't win against the respective champion as death interrupted the rivalry, but its team managed to "solve" the game.

Draw: Checkers

The history of machines playing Checkers is summarized by Kidwell (2015), he considers the 1868 AlAjeeb to be the first automaton to play Checkers, at the time those machines didn't use any sort of programming, instead an expert human player would hide and make the moves, and so Ajeeb would rarely lose.

In terms of AI the actual starting point was with Christopher Strachey and Arthur Lee Samuel who was influential in the field of AI, and one of the first to work on a game program with the goal of creating a

self-learning program, in a seminal paper Samuel (1959) that popularized the term "machine learning."

They both worked independently on a checkers program and Starchey published his work first in 1952 while Samuel had his program up and running in 1954, what his program stands out for was that it was designed to learn. (Samuel, 1959)

This program is cited as the achievement that Samuel is most famous for and what he labored the most into, which was 20 years according to Samuel himself. Weiss considered this to be the "world's first learning computer program" and the "first functioning artificial intelligence program". (Weiss, 1992, p. 69)

A Checkers board has 32 squares, the program assigns 1 to the bit assigned for each square with a piece on it, then looks ahead through a linear polynomial search tree, and as many other early game designs tries to, it gets rewards that push it to the ultimate reward of winning the game. The reward is a score system with weighted scales that are optimized with self-learning, the number of moves that it looks into is limited by some conditions to save computing time, conditions such as whether the next or last moves were a jump or if an exchange is possible, this is tested on plies to a maximum of 11 where it stops regardless. (Samuel, 1959)

Each board position in the tree of possible positions is evaluated through minmax decision making, this means that the program looks for the maximum score route and counter the opponent attempts to choose the route with the minimum score for the program. After choosing such a route it selects the move that leads there, then after the opponent makes their move, the program reassess the situation. To face the technical limitations of his time Samuel used a system of cataloging that removes redundancies and inferior positions. (Samuel, 1959)

The program didn't fare well playing against expert human opponents, even though it took only 8 to 10 hours of learning until it was able to beat the programmer himself, here we may note that a short T.A.G stages were completed if Samuel intended to improve his play by learning from his computer.

In 1966 IBM hosted the world checkers championship between Walter Hellman and Derek Oldbury with the condition that they play against the program, it lost all 8 games but won IBM 15 points rise in stock.

Later it also lost to another Checkers program that was created by Eric Jensen and Tom Truscott with support from Dr. Alan Bierman at Duke University, in 1979 they wanted a shot at playing against Marion Tinsley, the world Checkers champion who's nicknamed Terrible Tinsley in reverence, but they couldn't meet a bet of 5000$ that was proposed by the Checkers federation. Tinsley commented after seeing 6 games of their machine that it was an amateur, stating "Perhaps someday the programmers will have a real breakthrough. But until then let them behave like true scientists and refrain from undue boasting about their offspring". (Schaeffer, 1997, p. 99).

The breakthrough came within 20 years with the famous Checkers program Chinook, it diverged from the focus on self-learning and filled the gap between Samuel's program and AlphaZero. There were other Checkers programs, such as Gil Dodgen's checkers program, Adrian Millett's Sage Draughts, and Martin Bryant's Colossus which was the biggest rival to Chinook. Indeed, Colossus Draughts was the first checkers engine to win a human tournament in 1990, Tinsley can see that the programs were developing, and Chinook ran the gauntlet, so Tinsley accepted the challenge.

The main programmer of Chinook, Jonathan Schaeffer, started playing Chess at a young age, he later developed a Chess engine called Phoenix that was successful enough to tie for the first place in the World Computer Championship in 1986, the scientific community including Schaeffer focused on chess partly due to a misconception that Samuel has already solved Checkers.

A simple conversation pointed Schaeffer in the direction of Checkers, and one day after the 1989 Chess championship he made up his mind to redirect his efforts, until 1996 he worked with a team that will involve at different stages the following contributors: Norman Treloar, Joseph Culberson,

Brent Knight, Paul Lu and Duane Szafron. The program they designed would become the World Champion after a resignation from Marion Tinsley and a win against grandmaster Don Lafferty and it remained undefeated till its "retirement".

The details of this project are laid out in a book he published in 1997 titled "One Jump Ahead: Challenging Human Supremacy in Checkers". The following paragraphs are largely based on the book to highlight the Chinook journey.

Schaeffer started the design with the help of Norman Treloar who is a Checkers player, Schaeffer handled the programming and Treloar the expertise.

The evaluation function of the program, the function that assigns weights to positions according to multiple features like mobility, would undergo massive changes over the years. The point is to get the right pieces of knowledge into the program and assign the accurate weights to each so the program can logically evaluate the best possible position. It was handcrafted through trial and error, through playing games against the program and adjusting the weights accordingly. For the purposes of playing tournament matches under time control the function was of two kinds, quick and complete, the quick evaluation determines if a complete evaluation is required or not.

The manual change in evaluation is different from TD-Gammon since the human programmer is involved, the evaluation function would also be improved with each tournament the program participated in, by analyzing the results, troubleshooting mistakes, and countering some expert traps. Even subtle changes such as assigning a different number to the state of a draw since some draws have a better chance of breaking into a win than others.

The addition of human expertise is what makes Chinook different from Samuel's program, Samuel wanted the program to learn by itself while Schaeffer didn't see a reason why the program couldn't rely on existing knowledge as human players do. Schaeffer and his team benefited from human knowledge in many ways, one good example is the use of previous games in the opening book, the opening book didn't rely entirely on human knowledge because strong opponents may set traps during famous lines of play.

On the other side of the board was the endgame databases, the endgame in Checkers could have different number of pieces on each side, for each number there are millions of positions, for example the six-piece database has 2,503,611,964 positions, this momentous task was done by breaking down the positions to 4 disjoint problems and further down based on the number of kings on the board. More specific details about the evaluation function, the knowledge and the databases could be found in a paper titled "Reviving the game of checkers" by the Chinook team. (Schaeffer et al., 1991).

The program also required a lot of debugging and reprogramming at different points in time, debugging was a daily activity for months, the bugs included overlapping computations, repeated calculations, large offsets in evaluation function weights with overestimation or underestimation, software errors, some issues were inexplicable crashes at crucial times while playing against Tinsley.

Other issues could be considered as hardware shortcomings of the early nineties when computation power wasn't as advanced and storage was limited, it is almost astonishing to read statements in Schaeffer's book about the computer memory having 32 megabytes of RAM and that 460 megabytes was too large, yet it is also a testament to the ingenuity of the computer scientists that always have to deal with the limitations of their days despite knowing that in due time Moore's law would come for the rescue, after all it was a race to the top where Tinsley reigned as Champion that won every match in every tournament for 40 years except losses in less than 5 matches.

To give another example of this formidable human opponent, during the preparation for one of the matches against Tinsley, Chinook analyzed over 700 games and played similar games to look for the best moves, surprisingly Tinsley always picked the best move, even the few that were deemed wrong by Chinook's analysis would later prove to be right.

Chinook won first place in the 1989 Computer Olympiad and 1990 Mississippi State Championship, it was 2nd in the 1990 U.S. National Championship and the 1990 Computer Olympiad. The first human expert player that Chinook was able to win against was Ed Thompson, a former Canadian checkers champion, the informal game was played over the phone.

However, Chinook lost against Tinsley in their first encounter in 1990 and again in 1992 World Man-Machine Championship match, it also lost to Don Lafferty in 1991 and 1993, who was considered as 2nd to Tinsley. Later on the program improved enough to draw with both in 1994. Unfortunately Tinsley had to resign due to illness and the title was given to Chinook, a move contested by many, next year Chinook proved its worth by winning the World Man-Machine Championship against Lafferty.

While that marks the milestone of AI becoming on par with human grandmasters, it doesn't mark the end of Schaeffer's work on the game, in 1996 the 8-piece database was finalized, then starting in 2001 he continued to build the endgame databases and published a paper about solving the game of checkers (Schaeffer et al., 2007). The break was due to limitations in computer capabilities, Schaeffer states that the significant improvements allowed the computation of the 8-piece database in one month while previously it took 7 years. In 2005 the 10-piece databases were fully computed, with 39 trillion positions fitted into 237 gigabytes. This means that the database contains the result (win, draw, loss) of each of these positions.

The paper "Checkers is Solved" by Schaeffer et al. (2007) defines 3 ways a game may be "solved", first is ultraweakly which determines the result of a perfect play without knowing the strategy, the second is weakly which shows both the result and the strategy, this is the level at which Checkers was solved and it proved that perfect play leads to a draw. Lastly a strongly solved game is one where all possible positions are computed. One could argue that since Tinsley was by far the best human player and Chinook being the best Checkers computer then the draw result of their matches was already an indication of perfect play result.

Checkers was the first game of high complexity to be solved but wasn't the first game to be solved ever, in 1989 and a few years after many games were solved as well. Victor Allis and James Allen have independently discovered that Connect Four is always a win for the first player to move. The same was discovered for Go-Moku and Qubic, while Nine men's morris perfect play leads to a draw. (Schaeffer, 1997)

Finally the T.A.G in the story of Checkers has an overlap of the 2nd and 3rd stages, while the tool stage could be the Samuel's program, Chinook was definitely a worthy adversary of Tinsley, in fact Tinsley enjoyed playing against Chinook because it was daring enough in its moves, out of ignorance in some sense, as human opponents knew better than to risk it with a grandmaster like Tinsley; a computer treads where humans tremble.

The overlap between the adversary phase and the guide phase could be exemplified by what happened during a match between Tinsley and Elbert Lowder that was taking too much time, they decided to have Chinook adjudicate it and declare it a draw, the fact that two human champions could trust in the computer decision shows that it's more than a mere tool.

Defeat: Chess

When the topic of Man vs Machine is brought up it is common to think of movies like the Terminator hunting Sarah Connor or Kasparov losing to Deep Blue, the last section of the third chapter is dedicated to the former sense of malevolent AGI, in this section we will move to the famous chess match and review chess engines in general.

Chess is considered as one of the pinnacles of human intelligence, even before computers were able to handle a chess program, Alan Turing was already considering Chess as a benchmark for the machine thinking's ability (Turing, 1950), like Arthur Samuel he was interested in teaching the computer how to think, and they wanted a machine that can learn to the point where it can compete with man.

Turing was also one of the first to write a program for computer chess called Turochamp, it was ahead of its time and no computer was able to run it, but it could be considered as the first chess program.

AI and the computer industry owes much to Turing's vision (Muggleton, 2014), and perhaps his ideas about how machines think defined how programmers think about machines, and the goal of an adversarial position may have inadvertently given rise to the negative views that perceive AI as an adversary to humans forgetting that it is basically a tool told to play against humans.

The first tool that played chess was not a computer but an automaton like AlAjeeb, it was called the Mechanical Turk and had contraptions to move pieces on the board, a human player hid and fed the moves to it, it played Chess and won most players including Napoleon Bonaparte, but these tools didn't think for themselves.

Other early attempts that designed a chess playing machine was done by the Spanish engineer Leonardo Torres y Quevedo in 1912, called "el Ajedrecista", the machine had no human playing behind the scenes, but it was not able to play an entire game, it only played an end game variation with a king and a rook. (Williams, 2017)

The first paper to propose a machine that can play chess was by the founder of information theory, Claude Shannon (1950), who laid down some foundations to how a chess computer should operate, first showing the impossibility of doing a tree search that covers every possible move since a typical game may take 40 moves and each move has 10^3 possibilities which gives a staggering 10^{120} possibilities, the number of possibilities in Chess games are larger than even our modern computing powers could realistically handle. The answer is to use the minmax decision making, this approach took hold in game searches, and we saw it earlier used in Chinook for example. Another example of an idea used in other games is using different strategies for the 3 different phases of the game, as we saw already in BKG.

For the endgame phase, the concept of an endgame databases was popularized by Ken Thompson, who wrote the software for a game computer Belle, while Joe Condon worked on the hardware, Belle was the first computer to achieve master rating and the first to use specialized chess hardware, Belle won the ACM North American Computer Chess Championship 5 times and the 1980 World Computer Chess Championship. (Hsu, 2002)

Thompson made many contributions to the computer games community, according to Schaeffer a paper by Thompson (1982) on how strong chess game will become basically proved that the computer with the faster and deeper search will always win. That sets a clear path for programmers.

Thompson also dedicated an hour daily for three years to fill the opening book of Belle, the opening lines were from the 5-volumes Encyclopedia of Chess Openings, which amounts to 300 thousand moves.

The game community was not fragmented between the games, it was common to find the same person working or helping in multiple game software, both Schaeffer and Berliner were at some point working on their chess computers, Phoenix and Hitech respectively. Other notable chess computers include Cray Blitz designed by Robert Hyatt and Harry Nelson, more information about its architectural features could be find in Hyatt & Nelson (1990), Cray Blitz could search 200,000 positions per second, and it won the NA Computer Chess Championship two years in a row.

However, the computer that is most known to people is the one that managed to win against Kasparov, Deep Blue. Feng Hsiung Hsu who is one of the main programmers chronicled the events leading to the famous match in his book "Behind Deep Blue: Building the Computer That Defeated the World Chess Champion", this book is used as a main source for the following information.

Hsu started working on Chess computers 12 years before the match, he was still a student in Carnegie Mellon and was approached by the HiTech team, Hsu found issues with their approach to building their computer, the number of transistors used was too high and required a large circuit size if it was to

properly search for many plies ahead, he had a falling out with Mr. Berliner who was the head of the HiTech project and an animosity kept growing in the following years.

Instead of working with the team he started his own design on a computer he called Chiptest, starting with an examination of existing chess computers, some of them like Belle was already at a Master level in the game, he reduced the number of transistors by a factor of 150:1 with a priority decoder and shortened the length of the wires with a distributed arbiter circuit by a factor of 384:48.

Chiptest, like any chess computer had 3 main components: A move generator to find the chess move, an evaluation function to assess the quality of the positions ahead, and search control for the analysis of move sequences examined. The work on Chiptest was done rapidly, Hsu designed his own IO pads, and opted for brute search instead of selective searching which Berliner preferred, later down the line the favored approach was selective deepining which searches interesting moves further, akin to the quick and complete evaluation technique used by Chinook.

Soon Chiptest started playing and winning in tournaments against other strong computers such as Cray Blitz.

Hsu didn't work alone, for starters he used a tool created by one member working in HiTech, the tool compares "netlists" for different layouts in the move generator, this way the designer can verify that the layout matches the design, and it was important for the final verification phase. He worked with Murray Campbell in Carnegie Mellon, both later joined IBM to work on Chiptest's successor Deep Thought with Jerry Brody, A. Joseph Hoane Jr and Chung-Jen Tan, they also got help from Ken Thompson and from many Chess players.

In 1988 Deep Thought had better features, the project budget was only $5,000 dollars but it was beating much more expensive computers, the features included a hardware evaluation function that can recognize dynamic positional features, and automatic tuning of the evaluation function, this is done by comparing the computer moves to human master moves, Schaeffer burrowed and reprogrammed to work for Checkers but it proved to be less useful for Chinook.

Shortly thereafter it was competing with human opponents with impressive results, in the Software Toolworks championship, Brent Larsen was the first grandmaster to be defeated by a computer in tournament conditions, the program was also the co-winner with Grandmaster Anthony Miles ahead of 6 grandmasters making it the first time a computer finishes ahead of Grandmasters, earlier that year Kasparov was claiming that computers can't defeat grandmasters.

The next year an exhibition match was won by Kasparov, the version that played the match had a bug that didn't reward castling, and before the match Kasparov asked for and was granted the public games of Deep Thought, he prepared sequences which won him the second match. Hsu argues that they paid the price for this openness in sharing the games. However, the team wasn't expecting much against Kasparov this early into the game.

Deep Thought II took over, and playing against humans proved to help a lot in the upgrades of the sequence of computers that would end with Deep Blue, since certain bugs would only show during play, in a match against Bent Larsen the team realized the computer did not understand that it should trade pawns to empty diagonals for bishops. After a tournament in Hong Kong, Hsu believed in the need for a hardware repetition detector, as with Chinook there was also a lot of tuning in the evaluation function, another bug happened due to the special en passant move, a special move that haunts not only beginner Chess players but Hsu who spent a total of 6 months to fix problems related to this move across all chess engines he worked on. Another bug was that the program generated phantom queens, and some issues were out of their hands, in one of the matches a storm knocked the power in the computer lab forcing them to resign.

While most people know about the famous match in 1997, a lesser-known fact is that this wasn't the first match between the Deep Blue team and Kasparov under tournament conditions, one year prior Kasparov played and won against a predecessor, Deep Blue Jr. The win wasn't a clean sweep, Deep

Blue Jr snatched a win, before the rematch the team did major changes and testing, a new opening book they called "extended book" was used, it tries to capture the human concept of opening theory based on moves played by grandmasters in the same position.

Another upgrade was creating sets of tools to troubleshoot between games, Hsu makes the comparison that it's like pit-stops for cars in races. He also decided to create a new chip, a new repetition detector and new evaluation function and the computer required a new software model and a new move generator of adequate complexity.

Not only did Deep Blue learn from previous knowledge, but there was also active help by Grandmaster Joel Benjamin by playing daily chess against Deep Blue Jr, instead of playing full games he would retrace his steps and try new sequences. This helped the team find positions that caused problems for Deep Blue Jr. Another grandmaster, Miguel Illescas, was also invited to help with opening preparations, so were Nick DeFirmian and John Fedorowicz, there were also training matches against Grandmaster Larry Christiansen and Michael Rohde, both experienced in playing against computers.

Finally, the team was ready enough, Deep Blue was 5-10 faster than Deep Blue Jr. In theory the maximum search speed was one billion positions per second, the actual speed was 200 million positions per second. In the 1997 rematch Deep Blue won 2 times, Kasparov won once, and three games resulted in draws. It should be noted that Kasparov instead of trying to play like he normally would against a human opponent, he chose to play what is known as anti-computer chess he played seemingly sub optimal positions to confuse the program, it worked well against commercial computers at the time, he used Mieses Opening but this tactic backfired, and Deep Blue managed to win.

It's worth knowing that Kasparov never conceded defeat before this match. AI History was made.

Aftermath:

The computer that ran Deep Blue was dismantled right after the match, that didn't stop other chess engines from developing and advancing, the strongest nowadays is Stockfish according to Computer Chess Rating Lists website (CCRL 40/15 - Index)].

The win against Kasparov has been itched in our collective consciousness, it's still more famous than previous and even later wins of machine against man, at the same year a similar win took place but is far less discussed, it was a match between Logistello, an Othello computer, and Takeshi Murakami who was the champion at the time.

The game of chess as a whole wasn't negatively affected, it still enjoys worldwide popularity, in 2020 a mini-series titled Queen Gambit which was adapted from a novel about a fictional female chess grandmaster became the number one show on Netflix (Watercutter, 2020) and humans still hold championships between human grand masters with lofty prizes. The 2023 championship match in Astana, Kazakhstan between Ding Liren with Elo rating of 2811 and Ian Nepomniachtchi with a rating of 2793 had a €2 million prize. The game had a tense tie and Ding Liren won through tie breakers and received €1.1 million euros (Rodgers, 2023), the allure of human competition didn't subside even after computers became untouchable in the game.

The Deep Blue win did prove the superiority of computers and the T.A.G stages are the clearest with chess, for example the Deep Blue team was able to solve a classic chess problem that was 35 plies deep with only an 8 ply search using singular extensions (Schaeffer, 1997), this could be regarded as am early marker of the guide stage, on the other side Kasparov himself helped Frederic Friedel in creating a database for chess games called ChessBase, and he trained for the match against Deep Blue later by playing against commercial chess engines (Hsu, 2002). While the adversarial phase is clear in the arduous journey of Hsu, Campbell and the rest of the team to defeat the undisputed world champion, the guide stage started shortly after, with computer chess becoming a guide for up and coming players, human players now train by playing against the computer and studying its analysis unlike the old masters that trained only by reading human chess literature.

In fact, there is a clear distinction now between human moves and computer moves in chess, unfortunately this also allows for more cheating in online games and accusations of cheating in real life tournaments since the computing powers allow devices small enough to be hidden anywhere to have engines strong enough to win against grandmasters.

Another good marker of the guide stage is a match one year later between Kasparov and Veselin Topalov, in what is known as Centaur chess which allows a human player to have a computer assistant, this is a slightly old idea popularized by none other than Kasparov himself. A 6-game match was held in Spain between Kasparov with the aid of Fritz 5 and Topalov with the aid of ChessBase 7.0, the match ended in a 3-3 tie. They were only allowed to consult the databases for the 3rd and the 4th game and are allowed to use the analytical engines without the databases for the rest of the games.

Another variation is freestyle chess which allows players any form of consultation, in 2005 it didn't even take a grandmaster to win a tournament of this kind, two amateurs, Steven Cramton and Zackary Stephen, won such a tournament after having developed their own database (Baraniuk, 2022).

The year 2005 was the 2nd year in which a tournament of Man vs Machine World Team Championships was held in Bilbao, Spain, 3 world chess champions Alexander Khalifman, Ruslan Ponomariov and Rustam Kasimdzhanov lost 8 to 4 against chess computers Hydra, Junior and Fritz. One of the wins by Ponomariov is the last win by a human against a computer under normal tournament conditions (Chessbase, 2005).

Unlike Checkers, Chess is still not solved and may not be solvable for the foreseeable future due to its large number of positions, according to Schaeffer et al. (2007) Chess positions equal Checkers' squared and therefor won't be solved any time soon unless a new technology is developed.

In the meantime, the best any grandmaster can do is to attempt to play against a chess engine with a handicap, a handicap gives up pieces or moves to even the playing field, Paul Morphy who was considered a world champion in the 19th century refused to play chess unless he had a handicap, now the machines seem to have assumed this position.

Program Komodo played against grand master Hikaru Nakamura in 2016 with pawn and knight handicaps, Nakamura managed 3 draws before losing the final game (Copeland, 2016), for what it's worth humans can still manage a win with these odds, for example in 2020 Komodo lost to grand master David Smerdon 5 to 1 with knight odds in a rapid match (Doggers, 2020).

The last time a grandmaster tried to play against a computer chess without handicaps was in 2006, in a match between Vladimir Kramnik, who dethroned Kasparov, and Deep Fritz. 4 of the 6 games ended in a draw and the remaining two were won by Deep Fritz, computer science professor Monty Newborn McGill University, commented that "I don't know what one could get out of it at this point. The science is done" and that programming computers for competition purposes should turn its focus to poker and Go. (McClain, 2006).

In 2016, that is ten years after Newborn statement, an AI Go playing program managed to win against the champion and a year after that a Poker AI program won a tournament against the best poker players.

Before we go to Go, let's note that the centrality of chess in AI avenues may have come with multiple drawbacks, according to Ensmenger (2012) who questions the ramifications of chess becoming the equivalent of the fruit fly to genetics.

The paper starts off with citing the proclamations that chess is an intellectual game and thus a machine that plays it is intelligent, as we saw in the previous chapter, we can confirm that designing machines to play games was a concept that accompanied programming since the days of Babbage and tying it to machine thinking was set early on by Turing. Ensmenger (2012) argues that the steady success of chess engines also contributed to it becoming a mecca of AI research, however the focus on chess have overshadowed other domains and the algorithms used to win chess. Minmax and brute

force search became dominant which may have discouraged computer scientists from pursuing approaches, and how chess went from an attempt to simulate human thinking to an attempt at winning tournaments.

It may be true that chess research has narrowed AI focus for a while and made it all about certain techniques, but that couldn't be considered a heavy drawback in my opinion, if any other game was selected then it too will set a competition between techniques required to solve it and the focus will eventually be on the most successful technique. As for the aim of winning tournaments being more important than simulating thinking then that too is a natural benchmark of excellence in any game, the problem isn't with the focus on winning in the game but on the underlying argument that winning such a game is an indicator of thinking, in my opinion it only demonstrates the ability to play.

The tag of AI on Deep Blue is questioned by some, such as Makridakis (2017) that points out that it wasn't a self-learning program, Kaplan and Haenlein (2019) also consider it an expert system not an AI according to their definition. While it is true expert systems don't learn by themselves, they were designed to help in decision making, and according to the definition this thesis goes by, they are intelligent machines. Even if we accept that expert systems are not strictly AI, we will see in the next section that it was a bridging step till self-learning programs emerged within 20 years, and if it wasn't for the intelligent display of an "expert system" improving quickly and winning the highest accolades in strategic games then we may have considered games to be a dead end for a while.

Double Tap: Go

The game of Go doesn't have different pieces like chess and is played on a larger board than Checkers, an ancient strategic game on a 19x19 board favored by generals and intellectuals, the game is thousands of years old but its complexity and the location where it's most popular may have contributed to the fact that there are no automatons like The Turk or AlAjeeb that played it.

Its complexity was considered by Albert L. Zobrist who made the first attempt at a Go program in the late 60s, according to Zobrist (1969) a rough estimate of the moves possible in the moves tree is far larger for Go than for both Checkers and Chess, for Go to become as simple as Checkers it has to be played on a 6x6 board, and to match Chess the board would be 9x9.

Since the number of moves is staggering the approach must be different, it may also by the reason why programmers left this task to the end of the AI march against human grandmasters, a march that started with Zobrist (1969) who designed a program with two main parts, the first one transforms the board into an internal representation for the computer and saves the features in arrays, the second reads the representation and does calculations accordingly, his program was a simple first step that only managed to beat players who have less than 20 games experience in the game and no theory.

More programs followed such as Interim.2 by Walter Reitman and Bruce Wilcox in the seventies, The Many Faces of Go by David Fotland and Go++ by Michael Reiss in the eighties, and Handtalk by Chen Zhixing in the nineties, Go Intellect by Ken Chen, open-source programs like GNU Go, Pachi and Fuego as well as Crazy Stone by Rémi Coulom and Zen by Yoji Ojima in the 2000s, however the idea of facing a grand master player and winning was still inconceivable, so the programs would compete with each other in computer tournaments such as The Computer Olympiad and the annual Computer Go UEC Cup held at the University of Electro-Communications in Tokyo, Japan. Even the expectation of beating a junior human champion was still unreachable by the mid-2000s and a 1.5 million prize allocated for that achievement expired. (Cirasella and Kopec, 2006)

The gap between human Go players was bridged and surpassed when the crowning jewel of AI programs AlphaGo stepped onto the scene. It was developed by Google's DeepMind Technologies, according to its team Silver et al. (2016) the program learns to play through a Monte Carlo tree search and deep neural networks, it has three stages of machine learning, the first takes the simple representation of the board as input into a 13-layer policy network to predict expert moves from a

wealth of previous expert games, the second stages improves the policy network through gradient reinforcement learning, then the third stage tunes the position evaluation.

To evaluate the strength of the program before challenging humans it was petted in an internal tournament by the team against other commercial computer programs, Open-Source programs, and variants of itself. It won 494 out of 495 games, even with a handicap it managed to win between 77% and 99% of the matches. The next step was to challenge a human player, in 2015 AlphaGo played against Fan Hui who is the winner for 3 consecutive years of European Go championship, AlphaGo won all 5 matches to become the first Go program that defeats a human champion. (Silver et al., 2016)

The game of Go ranking goes up to 9 dan, Fan Hui was a professional 2 Dan player but the top player with 9 dan ranking is Lee Sedol, who could be considered as Tinsley or Kasparov in the sense that he was the last human champion. Sedol faced AlphaGo in 2016, a 5-match game for the prize of 1 million dollars. He started with confidence, declaring that he will win all 5 matches or just lose 1 match, and he got the final score right, it was 4 to 1, but it was the other way around. (Kohs, 2017)

This win stands out from the previous wins mentioned above, the program didn't take years of debugging and dozens of tournaments to have its functions hand crafted, the use of neural networks and the computation powers in the 2000s played a role in its success. Unlike previous engines that had to be spoon-fed opening books and endgame databases, AlphaGo would learn all that through reinforcement learning and playing like masters do.

Much like the progress of AI in chess, the only hope for champions to defeat an AI is to have it play with a handicap, in 2019 Lee Sedol tried again and played against HanDol, a Korean Go AI program, with two pieces handicap, Sedol snatched one win out of three. Unlike Kasparov who went on to play centaur chess and continue playing against humans, Sedol announced his resignation, his comment on the match could be an epitaph on human grandmasters' attempts to win against AI:

"Even if I become the number one, there is an entity that cannot be defeated." (Cheong-mo, 2019).

The guide stage in T.A.G was already present in the adversarial stage, some of the moves played by AlphaGo like move 37 in the second match and many others were unique and presented a new way of strategizing (Kohs, 2017).

Even though it won against a human champion, AlphaGo was still a program that learned to play by mimicking expert experience, the next step was to realize the old goal of Arthur Samuel and Alan Turing decades ago when they started working on Checkers and Chess, that is to create a program that learns the game from scratch.

One of the reasons Chess, Checkers and Go were popular in the programming community was that they had an abundance of human experience to test against and use to teach the program, but this level of data may have few drawbacks, as we saw previously Ken Thompson had to manually insert data for years into his Belle computer opening book, therefor the benefits of having an AI learning without leaning on expert data bypasses the need for a data set. This spares programmers the tedious task of inserting data or the expense of buying a data set, it also opens the door for games that aren't as ancient or as prestigious as Chess, Checkers and Go.

What makes humans grand masters in any game is sheer intellect plus the arduous training and learning from previous geniuses at their respective games, a new version of AlphaGo by Silver et al. (2017b) has the advantage of playing games and analyzing way faster than humans, and therefor instead of learning from what humans slowly accumulated it can learn it on its own. Through neural networks of improved self-play reinforcement learning algorithms and Monte Carlo tree search, the program played against itself and after 29 million games it ended up having a much higher efficiency compared to a second neural network that used the same architecture but learned from human play.

This approach can skip the adversarial stage and combine the use of AI as a tool and a guide, as it ought to be in my opinion, AlphaGo Zero managed to learn the fundamental and draw novel strategies, it was able to beat other AlphaGo variations including the ones that won against Lee Sedol

and Fan Hui. AlphaGo Zero came out on top with a significant margin, which attests to the power of self-learning in a game environment.

Still, AlphaGo Zero was a program that was tuned to play the game of Go, the holy grail of AI is to have programs that can function in multiple domains, the team came up with AlphaZero, a program that uses a general purpose Monte Carlo tree search and deep convolutional neural network to self-learn other games such as Chess, Shogi and Go itself, this method gave it an upper hand in all games in relatively short time, for shogi it took AlphaZero 2 hours to surpass Elmo and in Go it needed 30 hours to outgun the version of AlphaGo that won against Lee Sedol. In chess, it outperformed Stockfish after just 4 hours (Silver et al., 2018) while Hsu's journey from initial Chiptest till the Deep Blue victory took 12 years.

AlphaZero learned everything after receiving the basic rules of the games, the team went a step further with MuZero, which was thrown into the fray without the rules. Through trial and error, it learned the rules then went on to become better than AlphaZero, it also learned to play 57 Atari games. (Friedel, 2019)

One year after Alpha Go defeated Sedol a program called Libratus defeated top human players in Poker, furthermore another game that is less strategic and more knowledge based was won by an AI engine, in 2011 Watson won at Jeopardy, a game that requires answering questions that may be intuitive to us but a challenge to a computer, the unique challenge was to understand human language and look for the most plausible answer, it did so with over a hundred algorithms that assessed different aspects of the question and to pile bits of evidence until it's confident enough to give an answer. (IBM Corporation, 2011).

General remarks:

T.A.G Stages:

For a clearer view of the T.A.G stages you may refer to Appendix 1: T.A.G Stages Summary.

This chapter answers the first research question and the question in the title of this thesis, we saw how game engines became unbeatable. Now to focus on the sub research question, when it comes to human AI dynamics, the answer is the T.A.G stages that was introduced already, there are a few remarks that I believe are worthy of mention as well.

While the common view of all these matches is that they were between man and machine, this simply wasn't the case. Both Schaeffer and Hsu considered the matches to be between humans playing different roles, Schaeffer states that it was a match between men, or as Hsu puts it, the 1997 match wasn't between machine and man but rather between man as a performer and man as a toolmaker, indeed as the chapter shows there were immense amounts of human input and tinkering to beat Kasparov. Hsu also doesn't consider Deep Blue as intelligent; he described it as a tool that exhibits intelligent behavior.

The common conception about such achievements as machines evolving somehow autonomously and ominously is disproven by the long process that both the Deep Blue and Chinook teams as well as others put in their programs, the human programmers and experts spent many years tuning the computers evaluation functions and upgrading the opening books as with Ken Thompson spending and his opening book or the years it took to complete checkers endgame databases, to assign the win to the game engines is to ignore the human labor and ingenuity that was poured into them, and this clouds our judgement when it comes to ascribing the success to machines as separate entities.

It may be argued that no one literally believes those machines developed on their own, yet many of the worries about AI outsmarting us has the fast-paced evolution of such tools as one of its premises. That is why it must be put aside in any serious discussion. Another counterargument is to point out the self-learning engines like TD-Gammon, Logistello and AlphaZero managed to come along way without heavy human tinkering, this is a stronger argument and worthy of more consideration, but its strength in pointing out how powerful these tools are cannot jump to a conclusion of agency without another logical premise of a sentient spark of life, none of the self-learning engines chose to learn those games, they were programmed with the fundamental rules of each game and allowed to compute strategies all the way up to transcend human champions' abilities, even MuZero had to be fed data to learn, after that the machines didn't question their existence or show any sign of intelligence the way a human does. What I'm stating above may come across as obvious to some readers, however the discourse regarding AI is riddled with arguments that are founded on questioning what is stated.

Game engines did not become unbeatable because they chose to, it was entirely a human directive, effort, as well as hardware advances. In fact, the Tool – Adversary – Guide stages are not what programs naturally go through on their own, the history of game engines mostly shows how machine intelligence as a tool grows into a guide, the adversary is part of the game design, if humans had popular collaborative games, then AI advancements in those games would cross from Tool to Guide without any adversity.

When it comes to the hypothesized T.A.G idea, I'm not claiming that the second stage could be discarded, the focus here is proving that it's easy to confuse our use of the tool with something inherent in the tool itself, and if the hypothesis is true then it's important to know how it plays out in other AI applications. Take generative AI as an example, prior to it the digitalization of arts with software like Photoshop was a huge boon for artists, the fact that AI can now generate art with a text prompt and no skills caused a justifiable uproar in the online art communities, it exploits art without permission and raises questions regarding copyrights. This can be explained by entering the Adversary stage, then again, this adversary is not inherent in the tool, it was the use of this tool by humans that caused this uproar, and the tool itself didn't study color theory or aesthetic philosophy nor did it channel a deep subjective experience, it merely copies from artists who did that.

Therefor instead of being worried about the adversarial stage as if it's autonomous, any solution to a problem caused by this stage is a solution that could be asked from or enforced on the users of this technology. Politicians and managers should aim to pass through the adversarial stage to the guide stage with the least amount of human financial loss or any sort of loss.

Champions POV:

Another insight we can draw from reviewing the history of game engines is about several instances of human instincts that kicked in during matches, first example from a match between Brent Larsen and Deep Thought, the program seemed to have offered a free pawn by mistake, everyone thought it was the right move to take it, but Larsen declined, and later analysis proved that his intuition was correct.

Another example was during a game with Xie Jun, the first Asian female Chess Grandmaster. During the match there was an excellent move for her according to the computer analysis, but she missed it. Hsu noticed that she slowed down as if she sensed it. The program crashed and when Hsu restarted the machine, he accidentally wiped out the internal score file. Then she agreed to a rematch and when they reached the critical point, she paused again but didn't catch the move. After the match she mentioned that she must have missed a move and Hsu confirmed. On another board in one match between Tinsley and a human opponent, Tinsley "felt in his bones" that there was a draw if he picked a

move so he picked another and won when his opponent made a mistake, his gut feeling was confirmed accurate by Chinook later.

Neuroscientists and any researchers in any discipline interested in understanding human intellect and heuristics would do well to account for such hunches, it could be investigated by studying the brain while being engaged in such activities. But what is intuition? If we consider it to be something that happens at a very deep level in our minds in such a way that we may not describe the process of intuiting yet be more or less sure of the output, then the same could very well apply to the neural network processing in AI, in fact it is closer to how black box AI yields outputs than intelligent thinking since intelligence is more comprehensive and conscious than all narrow AI systems.

2 points may be concluded from the 3 instances mentioned above, Firstly, the extent of human intelligence is far beyond intelligible to us at the moment, and while science promises to solve everything it would be an article of faith to believe that we will come to completely understand it let alone create something that possesses it. Secondly without even reaching that far in understanding human cognition and problem solving we can already feel more, not less, affinity with AI when we think of intuition rather than intelligence, based on the literature reviewed for this thesis, an argument could be made that it's more accurate to consider AI as Artificial Intuition.

It is also worth mentioning that the human champions didn't have negative sentiments against game engines and databases all the way, we discussed how Kasparov helped build Chessbase and used it, as for Tinsley who was a very religious Christian, his view of the adversity against Chinook was inspired by his faith, he believed that he represented a player created by God whereas Chinook was created by man, Schaeffer quoted Dominic Lawson writing in the Financial Times who commented on this by saying that the opponent of Tinsley is not a soulless computer, but Jonathan Schaeffer himself who is equally a creation of God and whose instrument represents human ingenuity.

This shouldn't give the impression that Tinsley was dogmatic in his views of Chinook or of the programmer behind it, he enjoyed playing against Chinook, and he had a dream before one match, where God spoke to him and said that he liked Schaeffer too (Madrigal, 2017). Thus, a few words on the intersection of AI and religion are in order.

Guides to Gods: Apotheosized Intelligence

In general, there isn't much religious debate when it comes to AI, even though Turing anticipated a theological objection to the idea that machines can think, the objection revolves around an association between thinking and having a soul, and thus a soulless computer cannot think. He tried to address it by stating that in theological terms God creates souls and his power doesn't stop when it comes to granting souls to machines, and that we may be instruments in bringing this about by comparing it to procreation.

There is still no significant religious push back when it comes to AI as an entire field, other than on questionable ethical cases such as autonomous weaponry, in 2014 over 70 religious leaders signed a declaration asking for a preemptive ban on this technology (Religious Leaders Call for a Ban on Killer Robots – PAX, 2014). A decade later not much has changed or was implemented, a statement was recently made by Lieutenant General Richard G. Moore Jr. who serves as deputy chief of staff for plans and programs of the U.S. Air Force, he stated the ethics of the US army and the US society have Judeo-Christian roots, and this will guide the AI usage in warfare, he warned against countries - without naming names- that supposedly will not have the same ethical qualms (Avi-Yonah, 2023).

As for AI and religion there is no reason to believe that vague statements will remain so, more rigorous religious framing is required from world religions, AI is already transforming our daily lives

and that will raise question marks on the ethicality of using certain application. It is reasonable to predict that the human AI or AGI dynamics will involve more religious transformations, and perhaps even a revision of many religious underpinnings of the uniqueness of humanity. In fact, there was already 1 failed attempt to open an AI church. (Korosec, 2021b)

AI or robots might be worshiped as gods or AI might generate a new religion; Neil McArthur in an unpublished paper argues that the later will happen. Religions based on LLMs in chatbots like ChatGPT which give surprisingly great answers on a wide range of topics, this according to McArthur would inspire awe among users and will be convincing enough to merit worship when it's coupled with other AI features as immortality. The chatbot answers will elevate to doctrine, he then goes on to describe the characteristics of such religion(s) and even further into the future to discuss issues such as sectarianism, tax breaks and acceptance of the new religion (McArthur, pre-print).

The statements and arguments in the unpublished paper are more in line with speculative fiction than anything scientific, for starters the paper doesn't draw any direct analogy of how religions actually started, many things do inspire awe but the dedication in religious groups is much more than feelings, and it's too early for far reaching conclusions, early chatbots in the 60s such as ELIZA also stirred the emotions of some users, and indeed chatbots could give some vulnerable users comfort and consequently distress if the programming changed, but the emergence of a religion based on chatbot answers is an overkill for now. It's not impossible in the distant future when AI is more integrated and powerful enough to be a Guide in most or all fields.

A more realistic AI related quasi religion already exists, transhumanists believe that with technology it would be possible to achieve immortality, this is a promise that many religions offer in different shapes, the Elixir of Life in or the Pill of Immortality were both sought after items, the idea that we can upload our minds and live forever in servers could prove to be equally an imaginary quest that would motivate the brightest of men to no avail. Transhumanism also has a prophecy of the singularity, the concept of ASI fits neatly with this prophecy, believing in it seems to involve more faith than science, after all an ASI may not improve itself simply because it can, but the believers in such tales ascribe omnipotence and omniscience as necessary attributes like one would do to a god.

If the T.A.G is applied to the theological aspect then we can predict questions for each phase, as a tool what and how they are to be used in a way in line with the teaching of each respective religion, then as an adversary what are the ethical conditions of such adversity, and perhaps finally as a guide it could be invoked to settle religious debates or release fatwas. Metaphysical questions will be seriously pondered upon if an entity that is smarter than us was ever found or created. Another path that requires the question "Can machines think?" may very well become "Can machines believe?".

That might become a research question in the future, for the purposes of this thesis however, in the next chapter we will avert our gaze from the possibility of an AI heaven above and look below into the abyss; into the realistic dangers, and into what some people believe in already, an AGI hell.

Chapter 3: AI is king, now what? The potential risks of using AI.

The review of literature related to the risks that AI already poses is tricky, not only because it's evolving at an increasingly rapid pace but also because it's a contentious topic and according to some mounts up to a possible existential danger.

The aim of this chapter is twofold, the first is to present a proper framing that to make up the gap in literature and to describe an emergent pattern, the second is an attempt to classify multiple existing and potential risks into broad categories so that the answers to these risks could be dealt with in a systematic way. All while building on the information in the previous chapters to draw a holistic picture of the situation at the time of writing the thesis.

There are different ways of modeling the risks of AI. The EU Artificial Intelligence Act (AIA) ranks the risks according to severity in 4 categories from minimal to unacceptable and assigns obligations depending on the severity, Novelli et al. (2023) argue that the model fails to properly regulate AI due to its static nature and propose a new model that borrows from climate change risk models, it views the risks as consequence of Hazard, Exposure and Vulnerability.

This chapter is not based on this proposed model but doesn't contradict it necessarily, I propose a bird eye's view of the risks, filing them in 4 different categories with a focus on causality and composition rather than on threat levels and drivers. The first two categories are more meta in their nature, they include the risks of predictions and the risks that stem from AI as a technology like any other technology. The 3rd and 4th categories are more direct and are usually what comes to mind when the topic is brought up, the distinction between them is chronological. The 3rd category is present or pressing risks such as job loss, biases, military use. The 4th category is the speculative risks of when/if AI achieves or surpasses human levels of intelligence in a general sense.

Although there might be an overlap between some of the categories; a failure to predict may cause harm in the future and unlock a speculated risk, but I don't think any category could be done away with completely or collapsed into a subcategory. For example, an attempt to simplify it into chronological risks which are the 3rd and the 4th categories causes an issue with the 1st category since predictions belong to both the present and future plans and outcomes. The only future exception of mutual exclusivity is that the 4th category could become a subcategory of the 1st if AGI is objectively impossible, which is impossible to tell currently.

The model that is used by AIA and the climate change risk model that is refitted for AI by Novelli et al. (2023) could be transposed on the model brought forth in this chapter but that requires more research and is outside the scope of this thesis.

Figure 1 at the end of the chapter better illustrates the connections between the risks that will be explained in the following pages.

Risk I: The perils of prediction in AI:

As pointed out in the first chapter, making predictions might come at a heavy cost in the field of AI (Haenlein & Kaplan, 2019; Garvey, 2018), this category will deal with the concept of humans making predictions regarding what AI can and cannot do, although another reading of it could mean the perils of what AI itself predicts, but for the sake of neatly organizing the categories the risks of predictions made by AI is shelved under the 3rd category when it comes to reliability and what it might cause in terms of economic setbacks.

Making predictions when it comes to technology is tricky, a mistaken prediction may become a joke as with the 1903 case of the NYT claiming that it would take millions of years to develop aircrafts (Anslow, 2022), or could lead to project cancellation as with the cases of the US military's Autonomous Land Vehicle project or DARPA's Speech Understanding Research program (Crevier, D., 1993)., While reviewing the literature in the previous chapters one of the things that stood out was how often predictions about AI development were wrong, they were not wrong in one specific direction, some predictions were too ambitious while others too pessimistic, the predictions were sometimes true as statements but wrong in terms of duration.

This feature of AI development is critical and must be kept in mind while discussing any risk, at the same time it could be viewed as a risk of its own. This section will start with examples of predictions gone wrong from previous chapters literature, then move to literature that is focused on predictability of AI development.

Starting with the history of AI in games (Cirasella and Kopec, 2006), an example of critical prediction could be found in 1990 when Zia Mahmood, a grand master in the game of Bridge, offered 1 million pounds to any program that could beat him, 8 years later he played an exhibition match against GIB, a program that won the Computer Bridge World Championship, and despite winning Zia withdrew his challenge.

On the other hand, an ambitious claim was made among other claims in 1957 by Herbert Simon who was one of the creators of the Logic Theorist as stated in the 1st chapter, he predicted that a computer would be world champion within 10 years, his prediction came true only it was off by 30 extra years.

What happened in almost 10 years, in 1968, was that David Levy, a Scottish chess champion, made a bet with John McCarthy (Baraniuk, 2022), that a computer chess wouldn't beat Levy within 10 years, and he laid down £500 which is the equivalent of more than £8,000 today to stake his claim. He won the bet after a series of matches across the years against multiple chess engines such as Chess 4.5 & Chess 4.7, Kaissa and MacHack.

Levy made another bet against the first engine that would beat him, Cray Blitz attempted but lost 4-0, the bet that amounted to $5,000 was claimed by none other than Deep Thought in 1989 after winning against Levy 4-0, 31 years after he made the first bet. (Hsu, 2002).

Moving on to literature specific to AI predictions, Armstrong et al. (2014) argued for the importance of assessing the reliability of predictions related to AI due to the transformative or possibly destructive effects of this technology. The paper first breaks down predictions in 4 categories: Timelines and outcome predictions, Scenarios, Plans and Issues and Meta statements, and used 5 case studies of famous predictions: They are the initial Dartmouth conference, Dreyfus's criticism of AI, Searle's Chinese Room paper, Kurzweil's predictions in the 'Age of Spiritual Machines', and Omohundro's AI Drives.

As for the prediction methods, they could be loosely categorized into 6 types: 1. Causal models such as predicting an outcome based on underlying known causes such as physical laws. 2. non-Causal model where the underlying causes are unknown, but extrapolation of trends could lead to an accurate prediction such as Moore's law. 3. Outside view which is an implicit non-causal view that infers from groups of observations a certain trend such as the trend of revolutions in human history. The 4th method is Expert Judgement, which is self-explanatory, but the authors of the paper argue that it's not bullet proof, and 5th method is non-Expert Judgement which cannot match expert judgment, the area of expertise is of course related; a person who is an expert in any other field cannot carry over his prediction powers to field of AI. Finally, the 6th method is Philosophical arguments, which are logical and could serve the development of AI or simply warn against impossibilities.

We instantly see where the previous examples fall amongst these methods, the gambits by David Levy and Zia Mahmoud are technically non-expert predictions from an outside view which is their expertise in the relative games.

To assess timeline predictions Armstrong et al. (2014) took the results from a database by Machine Intelligence Research Institute that has 257 AI predictions made between 1950 and 2012, when it comes to the date that will see the development of human level AI, in the studied predictions it was observed that predictions are more likely to set the date after 15-25 years, and there was not much difference between expert and non-expert judgement, which makes timeline predictions quite dubious.

As for the 1956 Dartmouth conference case many wrong predictions were made by experts, they proposed to have 10 scientists work on AI for two months to figure out how to make machines use language, form abstractions, and become self-improving, and the ambitions included achieving a high ranking in chess, the ability to make music and theorem proving. Armstrong et al. (2014) argued that scientists of the time who were the leading experts in the field were reasonable in their assumptions that the matters they wanted AI to tackle seemed to be solved easily then even if it was objectively wrong. Which further makes it difficult to accept the validity of the prediction just because it comes

from an expert, or to assess it in a timely manner, since everything will be answered in hindsight but that's hardly any comfort for the time being.

McCorduck (2004) provides an account of Herbert Simon who justified his conference predictions, he stated that they were an attempt to make concrete possibilities as plausible extrapolations of current computer abilities, he attributed the delays in achieving the claims to the lack of people working on solving those problems, but as we saw in the previous chapters it also has to do with the sheer amount of time it takes to finely tune the programs and the limitations by computation powers at the time, we now know that AlphaGo didn't require a larger team, it required better hardware and software.

As for Dreyfus case who published a scathing criticism of the optimism and the promises that the AI community made, Armstrong et al. (2014) argued that his criticisms were accurate given the computer capabilities of the time, for example he noted that human language is ambiguous for machines, and that the AI community has computational paradigms to explain human behavior but that doesn't hold true, McCorduck (2004) dedicated an entire chapter to his criticisms that are based on refutations of AI community assumptions, his arguments fall under the outside view, the metastatament and the philosophical argument, namely a phenomenological view, the details of which are outside the scope of this thesis.

Although he was right in predicting the obstacles that the AI scientists will run into shortly after their initial successes, he didn't account for their ability to think of new methods later, he was snubbed by the faculty members that he railed against. The main takeaway from this aspect of his prediction is that an outsider could very well make a solid case against the promises of AI scientists, but Armstrong et al. (2014) believe that it gets tricky because outsiders couldn't always be taken this seriously, and that the strength of his argument hinges on the philosophical side not the outsider angle.

Another main takeaway is not from his writings but rather from a chess match that he was talked into playing, him being an amateur player made him lose to an early chess engine with low ranking, this didn't change his views but was used to mock him with word plays around the fact that computers can't play chess but neither can he, or that a child may win against a computer in chess but that computer could beat Dreyfus.

After reviewing the above examples, I believe that their juxtaposition presents a good idea of the perils of predictions when it comes to AI. It may be that an expert opinion is wrong while a non-expert who focuses on the philosophical underpinnings may get it right, but it's important to be precise in what predictions got wrong, in the cases above it was mainly the timeline that was off by decades yet the ability to achieve a certain goal, like computers winning a chess championship or winning against a particular player, eventually came true. The risks of falsely predicting an outcome could lead to AI winters or personal embarrassment.

The remaining cases that the article deals with feed the same conclusion and may not warrant further review for the context of this thesis, for example Ray Kurzweil's predictions were assessed for accuracy and at best it could be said that his predictions are 42% true (27% true and 15% weakly true), even if we consider the 11% of assessments that were undecided, the score goes up to 53%. While the paper takes this to be somewhat of an impressive record of his model, it does in my opinion prove the opposite, the trickiness of accepting expert judgements as good predictors, after all a 53% accuracy is slightly above a flip of a coin.

The article (Armstrong et al., 2014) is useful but slightly outdated due to the fast-paced development of AI and that's why a look at more recent articles is required. One more recent example of utter failure is Henry Markram's 2009 claim that he would reverse engineer human brains with a supercomputer and in 2013, the European Commission awarded his initiative a 1-billion-euro grant (Yong, 2019) yet the ambitious project failed to meet a fraction of its stated purpose and Markram kept rebranding and pleading for investment for other projects.

IBM's Watson -mentioned in the previous chapter- won at Jeopardy and the next day IBM announced Watson's new objective as an AI doctor, in theory just as it was trained on reading Jeopardy! clues it will be able to gain insights from reading medical literature, but it failed to make sense of it and is still far behind human doctors' ability to do so. It was unable to read medical records due to missing data or ambiguity and lack of chronological order despite its phenomenal NLP skills, IBM's Watson "overpromised and underdelivered" according to an article by Strickland (2021)

This raises concerns about the transferability of any AI abilities to other fields even if they seemed of similar nature, Watson's natural language processing that allowed it to win in Jeopardy didn't make it a good doctor, I would argue that this also proves that care should be taken when it comes to view the success of AI in playing games as an indicator of its dominance qua dominance.

The article by Strickland (2021) quotes doctors who argue that a change in the medical sector standards when it comes to sharing proprietary and privately collected data then Watson can become more effective, for now it's slightly useful for an instant second opinion, or in other words, we can say Watson is still in its tool phase, indeed it is already having success in some fields such as genomes.

From this section in the thesis, we may conclude that it will be a matter of time and ingenuity before it crosses to the next stages of adversity, when doctors start complaining about possible job losses, or directly as a guide.

Other than the Watson's case and Markram's brain project failure, Funk and Smith (2021b) gave examples of other failed predictions that led to projects cancellations such as Google flu trends or the critical response of experts to the success claims of Google's protein folding AI. Fully self-driving cars is another goal that has not been fully reached and predictions to have it in the next 7 years by an MIT task force might be another example that fits in this section.

Forbes does its yearly AI predictions and assesses the accuracy of the past year's, in its 2022 assessment Toews (2022) reports that only three out of the ten previous assumptions were wrong, and two were quite right and half were right. Further independent research into the accuracy of their predictions may give some insights but 50% accuracy is still not something to go by. Finally, it should be noted, and it goes without saying, that the word prediction must also be precise, a prediction of AI abilities is not the same as a prediction that an AI company will make certain revenue as is the case with some Forbes predictions.

Summary:

This category could be summed in the following statement: *Don't bet against AI abilities to achieve a goal if it stepped over the initial threshold of doing that task, you may only bet against AI attempting to venture into an entirely new field. And don't set a timer on your predictions*.

A recent example would be art generating AI ability to draw hands, one of the quickest ways to spot an AI generated image of humans is that it has trouble recognizing the anatomy of hands, it frequently adds or subtracts fingers and disfigures the shape, that flaw shouldn't be taken for granted as an insolvable problem as it is already getting remedied (Verma, 2023).

One of the risks associated with failed predictions is that it yields AI winter, an argument made by Yann LeCun in a post cited by Ford (2016), in this category I presented some cases where the funding was cut when the goals weren't met in a timely manner.

My statement about predictions is not meant to be generalized without a philosophical framing of what is possible, after all Armstrong et al (2014) concluded that predictions based on philosophical thinking are more accurate than those of sociologists and computer scientists (Piesing, 2013) because what AI can achieve may not necessarily mean what computer scientists take it to mean. As presented in the first chapter with Turing test being equal to a true benchmark of intelligence. At the same time any bet against AI has the burden of providing a falsifiable statement, otherwise the detractors of AI

capabilities can ignore the successes ad infinitum while accidentally strengthening the predictions in the process.

In fact my statement regarding predictions itself and the idea of T.A.G stages could both be taken as predictions, they belong to a non-expert judgment and a meta statement methods according to the cited categorization, the same article claims that a proper approach to test predictions should be done through extracting its assumptions and presenting a test to disprove it, and people who want to predict could do well to be direct in their assumptions and to provide a test.

In that regard I would say that the notion of T.A.G stages could be inaccurate if any AI application could instantly become a guide in any field, in other words these stages have the underlying assumption that the first step is always under the control of human operators, even if that stage was short, down to a few keystrokes to give the command, because it would still require initial human operation. The T.A.G stages therefor assume that AGI will not be achieved in the common conceptualization of it as an application or an agent that matches human intelligence in multiple fields, and therefor is able to make its own decisions and act on them without any input from humans. Simply put, I don't argue that a more intelligent AI could not be developed, the history of games showed the gulf between an initial program drafted on paper and a self-learning program that wins in multiple games, the argument is rather that such an AI won't be autonomous, it requires training and an objective from humans.

T.A.G also has the caveat that the Adversary stage is optional, which means that AI is unique as a technology compared to all other sorts of technology in its ability to become a guide whereas all tools humanity created so far, no matter how impressive or useful, never became anything more than tools in themselves.

The distinction of AI as a special form of technology may lead to specific risks, but I claim that those risks which are unique to AI are more distant, and the pressing risks related to AI are risks tied to its Tool stage and must be dealt with accordingly. Before presenting a brief overview of such risks in the 3rd category, I first have to bring up what I observed as somewhat of a gap in the research, that is the questions regarding technology which are rarely presented as such and when they are, then no viable solution is offered.

Risk II: Anti-Tech Views & Actions, Questions Regarding Technology

This category will deal with two aspects that are anti-technology, first is the human reactions to technological advancements and the second is the philosophical views that are critical of unchecked technological progress.

In the previous section we saw how predictions might fail specially when it comes to setting a deadline, that doesn't stop researchers concerned with AI from tracing the possible paths to the next milestones, as a report by Clarke et al. (2022) did by creating a seemingly comprehensive model of transformative risks of AI, and how we may reach what they called High Level Machine Intelligence (HLMI), defined as a level at which the machine can perform or quickly learn almost all economically relevant information processing tasks.

The model attempts to encompass all possible scenarios that would lead to that level, including the possibility that it may require brain emulation or drawing analogies to the evolution of intelligence itself.

The authors are open to modifications of their model, the strengths of this model in my opinion lie in that it satisfies Armstrong's requirement of predictions presenting their assumptions clearly enough and by considering various differing assumptions at the same time. On the other hand, it has 2 weaknesses, the first one is in some of its quantitative approaches of estimating the timeline of the

HLMI evolution through analogies, by referencing other scientific breakthroughs and their contingency on scientific efforts, successful or otherwise. They add up the time it took pursuing nuclear energy, the development of the internet and a cure for cancer and divide the two out of three successes over the average number of research years. I would describe this as a mere veneer of mathematics, and it will add to the risks associated of predictions when such poor formulas are used in a large model and most likely presented to the public as a warning sign in the form of "AGI expected in N years"

The second weakness in the model is what brings us to the central point of this category, which is the myriad of assumptions that are taken for granted when it comes to technology. The model by Clarke et al. (2022) has the underlying assumption that we can steer technological innovation rationally ignoring the role that chance played in many scientific breakthroughs, and even if we agree for the sake of the argument with their predictions there is still the much harder task of guiding society to deal with the drastically transformative nature of HLMI or AGI.

As Kaczynski (2016) argues in the first chapter of his book, that successful prediction and management of a society's development is only possible when empirical evidence is abundant, citing examples from Rome's sumptuary laws failing to regulate consumption and stop the decay, from Italy's 9th century laws to limit oppression and exploitation, and more recent laws like the US prohibition era that benefited organized crime. The butterfly effect is also taken to argue that even if we come up with a tight system of calculation to predict then its precision could be offset by any small variance.

Of course, governance cannot leave everything to chance just because tight regulation and predictions may fail, but the point here is deeper than a call for a laissez faire approach when it comes to society or the development of AI.

There are many assumptions baked in the discussions of AI risks that accept without any critical thinking how normalized it is to keep disrupting society and markets with innovative technologies, the direct risks that will be brought up in the next category are often brought up without reflecting on how accepting previous technologies became ubiquitous had unintended ramifications that are still not harmonized with our lives.

The lack of proper reflection stems from multiple factors to be briefly discussed in this section, such factors merit research on their own, one of the factors is waving ideas critical of technological progress off hand, usually the term "Luddite" is used in a derogatory fashion to describe such ideas, or the reprehensible fatal actions of some of the originators of such ideas, like Theodore Kaczynski's.

However, in both cases there is more than meets the eye. Starting with the Luddites, the current use of the term obscures much of the original story of the rebels who were called as such, Sale (1996) tells the story of these rebels that were extremely secretive and organized and focused their attacks on the machines in the textile industry at the dawn of the Industrial Revolution.

The origin of the "Luddite" name is not clear however one story is of a boy called Ned Lud who was a knitter that had enough of his master's abuse and took a hammer to the knitting frame he worked on, the rebels rallied under the fictitious name of Ned Lud and carried out attacks between 1811 - 1813.

The luddites weren't a group of anti-technology primitivists, but rather a group of disenfranchised textile workers who saw their work from their cottages replaced by rapid industrialization in the form of factories and factory equipment that they cannot compete with. Their assault started with direct attacks on machines with special care not to hurt civilians nor factory owners, later the factory owners fought back which caused bloodshed and finally it escalated to assassination attempts, the rebels caused enough havoc in Britain and thousands of soldiers were deployed to snuff out their rebellion.

The rebellion was short lived, and the industrialists had the final say of that day, but its true cause was not an irrational hatred of technology, it was a reaction to the disruptive technologies of that time. Their actions proved to be largely unsuccessful, a discussion of their ethics is outside the scope of this thesis, but their actions proved wrong for pragmatic reasons as well.

And if AI is to become a guide according to the T.A.G stages, the Luddites likewise will become a guide for anyone who is anti-technology, and the risks associated with the unreasonable use of AI may produce the conditions of many disgruntled workers and others affected in ways unpredictable to us.

The current category of risks must therefore be understood in two ways, one is that technological advancements of such a caliber might spur violent reactions, and in the opposite direction we must not wait until such drastic disruptions take place before wondering how to fix it and how it came to be, it's better to genuinely understand the sentiment of those affected negatively by technology, AI being no exception.

The Luddites didn't write their ideas in a treatise, what we have of their extant writings are mostly letters they sent before their attacks and so on, but if we fast forward over a century later, we come to a more sophisticated one-man neo luddite militia who followed their violent footsteps but wrote his criticism of industrialization in a sophisticated and succinct publications.

Theodore Kaczynski is known as the Unabomber, unlike the original Luddites he didn't directly attack machines but indirectly targeted individuals, his reprehensible acts claimed the lives of 3 and injured 23, instead of soldiers roaming the streets it took a hive of FBI agents to investigate the bombings that he set off by mailing explosive to his targets, he was captured in 1996 and sentenced to 8 consecutive life sentences. In June of 2023, while this chapter was being written, he was found dead in his prison and his death was ruled a suicide. (Balsamo et al., 2023)

Like the original luddites his actions proved to be, at least in the short run, both ethically questionable and pragmatically unsuccessful. The main argument that his attacks brought attention to his manifesto titled "The Industrial Society and its Future" could be countered with the reality that it didn't incite the anti-tech revolution that he had hoped for, his ideas are tainted with the actions, and this caused more damage to his cause than if he used his intelligence and credentials as a respectable mathematics professor to spread the word.

The reference to such a complicated individual in the context of this chapter is to further supplement my argument that there are certain flaws in our blind acceptance of the rapidity of technological development and certain assumptions such as the levels of control we have over how technology develops. It does seem counter intuitive to think of any technology prior to AI as directing itself, but that is the idea of thinkers like Kaczynski, who without anthropomorphizing technology draws a comparison of humans incrementally relying on technology as an alcoholic with a barrel of wine in front of him, drunkenly convincing himself that he will only consume small amounts.

We may add another real-life development to enhance his argument regarding the assumptions of reliance on technology, a few years ago if you entered any restaurant, you would be handed a menu printed on paper, now you are expected to scan a QR code somewhere on or around the table. If we consider this simple reality we can see how many assumptions there are about the technological tools wielded by all citizens; to scan the QR code you must have a mobile phone that has a camera and software that can read the code, it is also assumed that either you will have internet access on your mobile or that the restaurant will provide it, and deeper assumptions such as having electricity to charge the phone. Most of these technologies were not even conceived of by the average citizen just two decades ago, the word internet itself had to be explained on TV programs, now it's normal to assume everyone has access to technologies including a WiFi or Mobile Data source, Mobile Phones, Mobile Phones with cameras, QR codes, Software that reads the QR codes. And the more we scan our surroundings the clearer we note how much of our lives have changed and how much technology is assumed to have become a necessity not a luxury. It's unimaginable that anyone doesn't have an email address, or consider the fact that this thesis is meant to be uploaded to University of Hasselt's website not to be handed in on paper.

Many of the technological issues that are directly brought and addressed, such as privacy issues, don't take into account society's reliance on technology. The direct solution that companies offer when it comes to controlling the permissions of apps on your own device doesn't mean much when you have no say in the permissions that people around you have, and it wouldn't make sense to control

permissions around you, which ultimately means that your privacy is no longer a matter of choice. Even if companies developed sophisticated techniques to ensure your privacy you will still be at the mercy of more technology.

Here we may briefly touch upon the essay "The Question Concerning Technology" by the phenomenological German philosopher Martin Heidegger, his concept of enframing as a way modern technology reveals and orders nature in a way previous technology didn't, he writes "The revealing that rules in modern technology is a challenging [Herausfordern], which puts to nature the unreasonable demand that it supply energy that can be extracted and stored as such. But does this not hold true for the old windmill as well? No. Its sails do indeed turn in the wind; they are left entirely to the wind's blowing. But the windmill does not unlock energy from the air currents in order to store it." (Heidegger, 1977, p.14)

Perhaps we can apply this to how the way machine intelligence gathers information about us as humans and we must reflect on the concept of Big Data. Humans now use applications that trace and process literally every step they take and the number of sleeping hours. There are calories counting applications and all sorts of communication are logged in an unprecedented way, since data itself became a product then technological thinking views man as a data-mine, and man views himself as an excel sheet and a dashboard of performance to be enhanced in order to fit in social media milieus.

When Kaczynski was writing his manifesto the Deep Blue Kasparov match hasn't even played yet, he diagnosed the industrial society at large, and warned of the unpredictable consequences of embracing new technologies. In one of his letters (Kaczynski, 2022) to David Skrbina, who is a philosophy professor at the University of Michigan and whose unique anti-tech ideas are also worthy of consideration, Kaczynski states that the nature of industrial scale technologies forces everyone to use them and limits freedom in unanticipated ways, cities are designed around highways and streets now in way that was not imagined when automobiles were first invented, today people are forced to commute using vehicles and the choice not to do so would only inconvenience them without affecting technological progress, I would add that same goes for any sort of commute, the freedom not to use transportation is for all pragmatic purposes no longer an option.

This may seem like a non-issue, why would it be rational to use less efficient technologies or no technologies when they seemingly make our lives easier and faster? Cars allow us to go faster and QR codes reduce paper waste, the anti-tech answer doesn't discount the good but points the cost. More technology means there is more dependance on it and this leads to issues of a different nature, for Kaczynski it takes away our freedom gradually and has other consequences that we weren't prepared for. Car accidents in Belgium in 2022 alone were over 37 thousands, harming around 46 thousand person, 3,400 of those were seriously injured and 540 persons died, none of which would have taken place without these machines. (Road Accidents | Statbel, 2023), this is just in Belgium, if we zoom out, we find that the number of accident-related fatalities in the EU for 2022 were 20,600 (European Commission - Press Corner, 2023).

Still, it could be argued that there are pros of using automobiles and to counter the number of accidents and injuries by pointing out how many lives were saved thanks to the delivery of life saving drugs through such vehicles and quickly delivering emergency cases to get proper aid via ambulances.

However, the back and forth between the pros and cons of any specific technology misses yet another side of the risks that come from over reliance on technology, for Kaczynski this reliance will gradually transform society and force humans to live in distress because it is reshaping society at a rapid rate without giving humans the time to adapt, it goes against what humanity evolved through and what instincts we have, indeed we can approach this point from the opposite direction of someone who is enthusiastic about how everything is turning out for humanity, in his book Rosling et al. (2018) argues that the world is in a much better shape than before and that we are distressed for the wrong reasons.

The reasons that Hans Rosling refers to turn out to be what he calls instincts, instincts that are holding us back from appreciating how good things really are. One striking example of those instincts is what he called the blame instinct, that we need to find a clear and simple reason for why something bad has happened and tend to believe that some human with bad intentions caused the misfortune. What is striking about this example is that its opposite was brought up by Jacques Ellul, a French philosopher who wrote multiple books that question the direction of technological societies, in an interview (Rerun Productions, 1996) he speaks of the possible scenario of an electrical dam that bursts and asks who is to be held responsible for an accident when many people have worked on its cause. Is it the fault of the geologists, the engineers, the construction workers or the politicians? Each has a specific fragmented task and function in their zoomed in perspective of work, the conclusion is that none of them may be held responsible. What Rosling refers to as a blame instinct that should be done away with goes against the issues of governance and responsibilities. Of course, when real accidents take place an investigation and due process may pick someone who was negligent or malicious, but in cases where the failure is purely technical, we are asked to accept a world that goes against our basic intuitions and instincts.

Then Kaczynski warns of more invasive technologies like genetic engineering that will be used to directly change our very being, and he projects that such changes will not be up for the individual to decide. This could be easily replaced with AI technologies that may get developed in small chips, tiny enough to be embedded in our brains. Will we really have an option to refuse if this technology is unleashed? And if the measures that were put in place proved to be poorly designed, or that technologies that were readily available turned out to be extremely harmful, then should we ignore our blame instincts as Rosling puts it, or should we look for who is responsible to improve the ways of governance and decision making? And more importantly, how can we address the question of technological takeover of individual freedoms when we are ignoring the warnings from so called luddites?

With AI the situation gets more complicated, the responsibility of any catastrophe that may be produced by AI lies with who exactly? The software companies or the experts that took their word for it? If there was an issue with the training data or any other technical issue that the software engineers working on the product were unaware of, and a self-driving car makes the decision to run over pedestrians because the data somehow informed its decision, should we simply accept these tragedies as a fact of life or be extremely cautious when we design these systems and accept the possibility that some technologies should simply be banned as an intellectual with an anti-tech views would suggest?

The obvious answer that we should be cautious is not readily accepted by all the scientists and businesses rushing to improve their models, these questions are pressing now and shouldn't be delayed until AI reaches the Guide stage of the T.A.G or the higher possible stage of AGI. Because at that moment AI may introduce solutions that run deeply against our instincts and the Roslings of the world will tell us to accept them.

If we keep in mind that Kaczynski wrote his works before AI became as powerful as it is today and way before the Covid pandemic, we can't but agree that our individual autonomy had to be given up for the sake of the collective, it wasn't an option to leaves our houses during lockdowns nor was it really an option not to get vaccinated, and even if these were necessary measures to reduce the spread of the virus we shouldn't deny how distressing those years were and how the pandemic itself wouldn't have spread across the world if it wasn't for the interconnectivity of industrial societies. The same interconnectivity undoubtedly allows us many services that make our lives easier than ever in human history.

Thus, the anti-tech views could be understood as criticisms of over glorification of technological progress, they present the dark side of the moon and the price we pay to maintain the technological societies we live in, a price that includes our psychological being, dignity and our freedom. The way out of a technological society would sound to us like the ramblings of a mad man who decided to live in the wilderness, but that shouldn't blind us to the fact that other philosophers raised questions about

the nature of technology and what technology does to our nature as humans or to the natural world around us.

Many like-minded thinkers existed between the luddites and the neo-luddites, Skrbina (2014), himself holds views that question the trajectory of technology, he lays down the ideas of many thinkers and his brief history of technological critiques could easily be transposed on the history of AI in the 1st chapter. While many of the thinkers' ideas are worthy of discussion at depth it would go outside the scope of this chapter, which is to highlight the risks related to AI, and the scope of this section is to challenge the idea that technology isn't an autonomous tool but a set of tools that humans have absolute control over. This wasn't the case with technology even before AI according to the likes of Kaczynski, Ellul and Skrbina among many others.

Without going into depths of all the ideas of the intellectuals Skrbina brings up, some of their points are worthy of mention, such as the one made by Georg Simmel about the mistake that we make by assigning the same benign and helpful values of each technological device on its own to technology as a whole, and how we put tremendous effort to develop and maintain the technological advancement in a way that may beat its purpose as something to control nature, we are instead enslaved by this new artificial nature. (Skrbina, 2014)

A.N. Whitehead is another thinker that made some observations about technology as a self-developing force and how human value is excluded in the more mechanized and competitive worldviews of his time, later on Oswald Spengler also commented on the unforeseeable consequences of using technology, him and others including Karl Jaspers lament the fact that workers in our societies turn into cogs without comprehending the full view of their work or deriving satisfaction from it (Skrbina, 2014).

AI enthusiasts usually sell the line that AI will allow people to stop working this unfulfilling jobs and have the time to pursue their own talents and hobbies, but as mentioned in previous chapters, the art generating AI is already threatening the notion of having a talent, soon those talents might evolve into how a person can command the AI and what prompts to use instead of learning any artistic skill, this to current artists is a complete mockery of their hard work. While users of art generating AI liken it to previous technological devices like the camera and how that didn't remove the art of painting but created new schools of it, other AI enthusiasts would argue that these tools are helpful even for the artists themselves. I personally think that the T.A.G stages apply with generative AI as well and that we are in an Adversity stage. If that is true the right move is to quickly ascend to the Guide stage, however that would still not address the risks mentioned in this section, an advanced art generating AI only means more reliance on technology.

Many of the critiques that Skrbina (2014) brings up share in common the view that technology is developing in an autonomous fashion with varying degrees of technological determinism, and these critiques have varying degrees of pessimism in countering the negative consequences of this development, they also seem to doubt the neutrality of technology and lean to focus on the negative sides of it without discounting the good, which runs against the common view of our times regarding technology; that it is a positive or a neutral force and the way it's used is completely up to humans.

Summary:

In my review of the literature of this category, which I would describe as anti-tech views, and as we saw with the luddites old and new, their main fault in my opinion was not in their theorizing. Most of the intellectuals bring forth compelling arguments about the transformative nature of technology, many proper points are raised if they were read with the charitable principle, but the fault is in their proposed solutions to the problems they accurately diagnose. Most thinkers didn't go as far as luddites or Kaczynski, but they don't offer a useful answer either, the answers revolve around a shift in our worldview without accounting for the human forces that drive technological advancement or a logical

defeatist conclusion of true belief in hard technological determinism, all of this further proves their initial point about the unstoppable tides of technology. And if we don't look for an answer that is not reckless vandalism or violence or intellectual appeals to a magical shift in how society conducts itself or nihilism, then we might as well wait until AI becomes a guide and blindly follow its instructions.

AI is no exception to the criticisms leveraged against technology, if anything some criticisms such as the autonomy of technology applies even more to it, so would be the lack of assigned responsibilities when accidents inevitably occur. In our attempts to wall ourselves in against the wrath of nature, as Ellul (1964) puts it in the 6th chapter of his book, we created a new artificial nature that insulates us.

The catastrophes that accompany the highly connected world that sustains this level of technological advancements will turn us into unwitting cogs, the use of AI will increase, and its effects will transform society and business alike. That is why we need a more thorough philosophical grounding of the values that AI should be trained with and answer the questions that were raised in the previous century before diving into a new revolution that is shaping up to be as transformative as the industrial revolution.

The risks in this category could be summarized as the following:

Those affected by the rising powers of AI will react in ways that might be violent, it's important therefore to make sure whatever impacts AI has will not drive anyone to their breaking point.

To do so we must listen to the valid criticism of technology as an entire system we have erected ourselves, and not limit ourselves to criticisms of this or that piece of technology independently, nor should we focus exclusively on the political or the economic models that may take some of the blame.

One of the issues raised by people with an Anti-Tech view is that technology limits our freedoms, it makes us rely on it and offers more technological solutions to the problems it induces, and if we advance too rapidly then most of the instincts we evolved to have will become useless and we will despair due to that mismatch. Some would say humans always adapt and this is another step of evolution, one possible retort is that adaptation to harsh surroundings doesn't necessarily mean becoming a better being, a prisoner serving a life sentence may adapt to prison life but that doesn't make him better than a free man.

The very neutrality of technology and AI must be questioned, and we must make sure to reign in technological development so that it serves us, not the other way around. Admittedly, this is easier said than done.

Risk III: Current and Concrete Risks

This category will be divided into 4 parts to briefly give an overview of risks that are the most direct and pressing. These risks are explained separately but they interlock to create harder problems.

1. Economic setbacks:

First let us present some counter arguments to the idea that AI will cause any long lasting or severe damage to the economy by turning to a substack article by Andreessen (2023b), the founder of Netscape and a venture capitalist who's a member of the US Homeland Security Advisory Council.

Andreessen attempts to address multiple concerns surrounding AI, the two risks that are relevant here are AI taking away jobs or Increasing Inequality, Andreessen believes that AI is another new technology that will not completely wipe out jobs, he invokes the Luddites, the Outsourcing panic and

Automation panic, how none of that caused mass unemployment, instead what we get with AI is productivity growth. He also brings up the notion of labor lump fallacy, which is the fallacy of believing that labor is finite, and so if AI takes away some jobs, then less so will be left for human workers.

For the inequality risk he states that it is in the interest of technology owners to sell it to customers not to hoard it by themselves and compares this concern to the Marxist claim that the bourgeois will steal all wealth from the proletariat. But this didn't turn out to be the case and by comparison it won't be the case for AI.

Now we may look at the other side of the arguments, as we saw in the previous section, the luddites were reacting to the way machines was taking their jobs and ruining the way of living that they were accustomed to, the acts started off with vandalism directed at machines but bloodshed wasn't avoided in the upcoming months, a century later the Unabomber had more sophisticated attacks against innocent individuals and a manifesto that aims to incite a revolution against technology as it stands. All of this was prior to the emergence of AI as a powerful tool capable of matching and surpassing humans in intellectual skills that were marked safe in previous technological advances. The point being is that Andreessen views the worries as irrational and theoretical, without addressing the direct damage it causes to replaced or deskilled workers, regardless of how well the economy does from a bird's eye view, and subsequently he ignores the damage that they themselves will also cause for society, perhaps rightfully so.

In his book "Rise of the Robots: Technology and The Threat of Jobless Future" Ford (2016) tracks some of the ways technology and AI will replace work and may cause different kinds of economical setbacks, this book will be the main source for this section and the following paragraphs extract the most relevant ideas.

Starting with automation technologies such as Kiva Systems used by Amazon starting in 2012, the Kiva robots autonomously use a grid and barcodes to deliver pallets and shelving units to workers, a decade later Amazon announced a new autonomous robot that is not restricted by the grid, as well as other robotics to automate many warehouse activities. (Wessling, 2022).

This is but one example of the type of work that might be replaced by robotics, as Ford (2016) argues in the following chapters that information technology as a whole is a uniquely disruptive force, in the 2nd chapter he argues that seven economic trend when considered together and adding technology to the mixture makes the disruption different than previous ones, such as stagnant wages, declining labor participation, and declining incomes and recent graduate unemployment, and polarization.

I'll refer to an example from the 7th trend and how it relates to technological advancement, Ford argues that after recessions the new jobs that are created are mainly in low-wage sectors, this means that simply finding work doesn't mean finding good work during recoveries and more importantly, in the case of unemployment caused by AI powered robotics and AI in general, some low skill sectors like fast food services that was a minimum wage haven might become obsolete.

Information technology disruption will be different for other reasons according to Ford (2016), its rapid advancement is unlike other technologies, Moore's law is one example of that, another unique thing about technology is self-replication which means that AI has a vast comparative advantage over human workers. Humans might choose to work with their best skills and rely on others for other skills even if they are better than them. To borrow an example from Ford, an excellent doctor who happens to be a good cook might choose to work full time as a doctor and rely on someone else to cook for them, AI is not limited in the same way because it can do both and without replication it may multitask at a much higher speed. This isn't the only issue when it comes to the special skills of technological replacement for human workers, deskilling of the jobs might occur due to partial automation, this allows businesses to have higher turnover rates and puts human workers at a disadvantage.

Another discontinuity in how disruptive AI could be lies in the fact that economies grow through demand and consumption not merely through supply and production, yet in the case of technological

alternatives to human employees, Ford argues that these robots will not be consumers of the goods they produce, they will only require power and maintenance to keep running the show. He points to the inequality in consumption where a small percentage of consumers have the lion share, and if businesses are geared to produce for the elite, then a sort of technological feudalism may arise, only this time the serfs will be programs and perhaps a few programmers too.

Another example Andreessen brings up is that Elon Musk became rich by making very affordable cars from the profits of selling less affordable cars, I do not think that citing the billionaire class which is the clearest example of inequality does his argument any service. I risk stating an argumentum ad hominem by pointing out that Andreessen is a billionaire himself. It's clear he's addressing technological inequality, which is a possible risk, but the term inequality usually refers to economic inequality.

During the time of writing this thesis, many impressive LLMs and generative AI applications were out in the market, their low price ranges supports Andreessen claim that such tech will be democratized, what it missed in reality is that this type of democracy of skills significantly reduced the barrier of entry into working as an artist that depend on online commissions, the years of hard work to perfect a skill and the outcomes of countless artists were put in a blender and people with no skills became competitors overnight.

I must state what is now becoming obvious but was not a few years ago: None of the current jobs -as we know them- are safe, therefor politicians, business owners and specialized workers shouldn't view this issue as a low skilled labor issue or a manual labor issue. It is understandable that AI enthusiasts will point to long gone jobs, such as bowling pinsetters, and state that current jobs should be no different, which might be the case but that can only be acceptable if it happens at a normal rate where workers are guaranteed to switch jobs with financially secure and dignified.

Finding solutions to these complicated problems would turn this to a PhD dissertation, but I would like to bring forth a simple idea that the balance is the key, a balance between the rational choice of using AI and the ethical choice of allowing humans to keep their jobs and not expect them to accept being rag dolled between jobs, skills and locations. One way is to set a replacement threshold, once technology proves its ability to effectively take over, a plan is set to slowly phase out human workers of that job. The current employees would be the last generation of that profession, allowed to keep their jobs, or more realistically the government intervenes, or a regulation should force businesses to compensate such workers if they are to be let go due to AI taking over, and further training or education for that job is immediately stopped.

2. The Alignment Problem:

In the 2nd chapter we saw how the game computers were trained gradually either by fine tuning or reinforcement learning, the final goal of winning the game is realized through crumbs of rewards for the program, the most simple and direct analogy of the alignment problem is to imagine a chess engine concluding that since winning becomes likelier if you have one more queen than the opponent then the best strategy is to advance all the pawns and promoting them to queens.

If the problem stopped at simple mistakes such as this then it would be a programming hiccup and not a risk, unfortunately as AI is given more complicated tasks that humans intuitively consider in a holistic way, it may become misaligned in ways that are hard to spot or solve.

Christian (2021) provides a thorough walkthrough of this risk in his book "The Alignment problem: How Can Artificial Intelligence Learn Human Values?", the book is based on almost 100 interviews and 4 years of research, the book definition of the problem is that AI values may diverge from its programmers' intentions, this definition fits in this category.

A stretched definition makes this problem a blanket term for the entire chapter, for example scientists didn't intend for the artificial surroundings that they created bit by bit to ease our lives to become a life support machine that we can't live without, nor was the risk of massive layouts due to AI advancements the end goal of any programmer working on self-driving vehicles. Even the last category that will deal with AGI and the possible extinction risk, that would be a clear divergence from what Hsu and Schaeffer were setting up when they designed computers that play chess and checkers.

Although there is much to be said about the moral responsibility of all scientists regarding their inventions, this philosophical discussion is more fitting in the 2nd category but is unfortunately outside the scope of the thesis either way. All of this is to state that the alignment problem in this section is taken in the narrow definition of divergence of values in a single AI system.

A few examples from Christian's book will be given to illustrate how the problem manifests in different contexts, the first comes from the risk assessment tool Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), the algorithm helps judges making decisions regarding bail, parole and the potential of recidivism by defendants on a 10-point scale. A few years after its use a scandal broke out when a group of journalists at ProPublica assessed the tools assessments and how reliable the results were in retrospect, they found unusual scoring that seemed racially biased.

Racial bias in algorithms is spotted in many applications, the book also relays the story of how Google Photos labelled a photo of friends as a photo of animals causing an uproar that Google fixed only by deactivating the label, an entire documentary titled Coded Bias by Kantayya (2020) tackles the issue of discrimination that results from skewed training data sets.

There is a thin silver lining to the issue of bias in code is that it may mirror the biases in society as a whole or the data set, and from there the biases could be acknowledged and dealt with.

The book (Christian, 2021) does a great job at presenting the alignment problem and there is few angles that the author missed except, in my opinion, one that could be in the blind side of any well intentioned discourse of this topic: AI may diverge from the values of its creators but human values diverge already. There is no set of universal values, only the most basic notions are global, like thou shalt not kill, but different cultures don't agree on anything beyond that even if they are perfectly okay with dealing with each other and coexisting in many contexts. While writing this thesis the 2022 Qatar Cup was held and the issue of the country's views on sexual orientations was deemed problematic to Western nations, the world is far from being unanimous on all ethical issues, civil liberties and values.

This may raise ethical concerns, some may distress the programmers working on solutions to be shipped to other cultures, indeed the split in the ethics of the programmer and the user is best represented in the story of how more than 4 thousand Google employees signed a petition to cancel an AI project for the Pentagon (Somerville, 2018), which brings us to the next set of pressing risks.

3. Military AI:

AI has already proved its steps in the Adversity stage in military settings, an article aptly titled ("AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis," 2020) announced that in August of 2020 an AI agent developed by Heron Systems defeated a seasoned F-16 fighter pilot, the dogfight was part of a series of simulated combat engagements as a part of DARPA's trials that were designed to expand a base of AI developers for its Air Combat Evolution program.

Of course, advanced armies are already aware of the risks that come with using AI, the 1st global Summit on Responsible Artificial Intelligence in the Military Domain (REAIM) was hosted by the government of the Netherlands on February 2023, the Summit included 2000 attendees from 100 countries and 80 government representatives contributing to the responsible use and deployment of

military AI. The 2-day summit resulted in a joint call to action on the responsible use of AI in the military domain.

The joint call to action acknowledges the potential impacts and the current lack of sufficient comprehension of the implications of its use, the call supports the idea of ensuring human responsibility and accountability and acknowledges the need to have a holistic approach with close attention to all stages of development. As the summit involved multiple stakeholders it asked for collaboration of everyone involved to make sure AI is used responsibly as it gets deployed in the military. The call to action also involved many points including the support of academia, knowledge institutes and think tanks globally to conduct additional research to better comprehend the impacts. (REAIM Call to Action, 2023).

The effort and the call to action with its 25 points are all good on paper, the call addresses the main risk of rushing AI use without fully understanding its impacts, one of the panels during the summit also brought up the biases that could be encoded in AI and how dangerous that may be in wars which are dirty and biased by definition.

Three years prior to this summit the United States' Department of Defense (DoD) released a list of 5 ethical principles of using AI, Responsible, Equitable, Traceable, Reliable and Governable. The Equitable principle addresses the alignment problem, and the Governable principle makes sure that it could be disengaged if need be. (U.S. Department of Defense, 2020).

The Traceable principle states that relevant personnel must possess an appropriate understanding of the technology, this principle could become shaky as we move on to the next set of risks.

It is good to see that the governments of advanced nations are not neglecting risks but other voices in the US military are impatient with the slow integration of AI since that will lead to a competitive disadvantage against China, the obvious benefits and the promise of powerful systems makes it hard for some to wait for the low-level application of high-level ideas or abstract ethical principles, another issue is bureaucracy.

Eric Schmidt, the former CEO and executive chairman of Google, opined on the outdated and the bureaucratically burdened state of technology in the US army, the solution to these problems was for the US army to tap into the private sector to speed up innovation. (Knight, 2023)

There are many uses for AI in warfare, a brief example could be provided by another private sector company Sentient Digital, the uses of AI include but are not limited to the following: Strategic decision making, Combat Simulation, Target Recognition, Threat Monitoring, Cybersecurity and Drone Swarms which apply swarm intelligence, that is the decentralized decision making where no single leader gives the order but the swarm achieves objectives as a collective. (Sentient Digital, Inc., 2023)

Schmidt believes that AI could be a game changer as impactful as nuclear weapons, whether his assessment turns out to be accurate or not, the mere perception of weaponized AI being this powerful will incentivize superpowers to have AI Manhattan Projects. On top of that we can anticipate the disastrous risk of combining such powerful weapons, consider the weapons of mass destruction are meant to be launched depending on readings from detection systems, or as first strikes, and if AI has a say, then it could be deployed if the systems give false warnings.

Such scenario isn't fictional, false readings in 1983 could have started a nuclear war and crisis was only averted because the human operator, the soviet Lt. Col. Stanislav Petrov, decided to trust his instincts and understandings of the short comings of the missile detection satellites, he decided that the system gave a false alarm of intercontinental ballistic missiles and reported accordingly, it took months for the experts to analyze what went wrong. (Air & Space Forces Magazine, 2023).

These scenarios are extremely dangerous but are by no means the only risks related to using AI in warfare, a report by Morgan et al. (2020) properly provides a taxonomy of such risks, they are organized in under 3 labels: Ethical and Legal, Operational and finally Strategic. It's outside the thesis scope to go into detail but a few examples could give a general idea of what's at stake.

An example from the first label is the Law of Armed Conflict that intends to minimize harms to civilians through distinction between them and combatants, autonomous weapons are still not good enough to make subtle distinctions specially in cases of guerrilla warfare. Another example is the question of who to hold accountable when it comes to autonomous weapons that may break the law. Even if those considerations were addressed there is an argument of human dignity, that it would always be wrong for a machine to take any human life. The question of responsibility sits in the previous category.

The Operational category includes risks that come from AI failing in unintended ways, this relates to how much trust we can put in the AI system and how reliable it is when it's deployed in real combat despite being trained in controlled environments. These issues will be better explained in the next category.

Finally, the Strategic label consists of the risks that will change the nature of warfare in general, an example is that the ease of deploying AI weapons without human operators might incentivize leaders to do so instead of pursuing nonmilitary resolutions. Flash wars as the Stanislav Petrov's scenario above also fall under this category, so is the risk of undermining strategic stability not through false readings but willingly as AI might become advanced enough to nullify the danger of hostile states to use their nuclear arsenals effectively.

The more AI gets integrated in manning weapons and operational or strategic decision making, the less control humans will have over their fates, the next risk about the level of trust we must put in AI outcomes will flesh out this point further.

4. Epistemic Regression:

Deepfakes (Deep Learning + Fake) are a type of digitally altered media that is recently becoming more and more powerful, what's dangerous about this type of digital manipulation is that it can apply the facial features of anyone on top of any face far better than previous digital manipulation tools. If we add the AI voice generators that can mimic voices to the mix then it is easy to see where things might go wrong, and already did, when a series of scam calls demanding ransom in return of kidnapped loved ones who were not kidnapped but whose voices were faked during phone calls (Karimi, 2023).

As AI becomes more powerful in reading the slightest changes in our gestures, facial features, tone of voice, and even to generate faces from voice clips incorporating the emotions and expressions (Zheng et al., 2021). All this will lead to videos that are indistinguishable to the naked human eye, this risk loops back into other risks when we consider that it warrants another technological solution in the shape of software that detects manipulation, soon we may no longer be able to trust our senses and will have to rely on programs to tell us what is real and what isn't. It also loops back to the issue of our reliance on the new artificial edifice we're setting up that takes us on a sharp turn in terms of our evolution as a species.

The educational system is another epistemic area already affected by LLMs, Milano et al. (2023) highlight the impact of these models on higher education and assessment of students, some compare using LLMs to plagiarism, but it is still hard to detect it at this point. As with the previous point, the technological tool of checking if an LLM was used, which loops back into the reduced agency of professors, while integrating the model into education threatens to loop back into the Alignment problem since these models carry their own bias with them. One possible way to avoid LLMs is to return to paper and pen assignments. This is not feasible and hardly applicable but the suggestion of it is a unique effect of AI as a tool that forces humans to go back to previous tools.

Since some may argue that this development is still acceptable since we are using more powerful tools that require powerful measures, and AI is ultimately a tool that helps us in decision-making. The issue with this acceptance is that it's not limited to discerning if we're watching an altered video or if a

student is cheating, AI could be used in more sensitive fields, as mentioned above it will concern military decisions, and as mentioned previously, it's set to be used in medicine.

While some like Rudin (2019) argued for more interpretable models in medicine, Coveney and Highfield (2021) call for more analogue options and less reliance on the digital, and for firmer theoretical underpinning instead of trusting computers gobbling data without minding the errors. Others (Durán & Jongsma, 2021) argue that we can trust black box AI without fully understanding the inner workings of their decision-making processes, they base this on the concept of Computational Reliabilism (CR), which is presented by Durán and Formanek (2018) as a possible answer to the issue of Epistemic Opacity (EO).

Simply put, a process is considered to be epistemically opaque (EO) if the epistemically relevant elements of the process are unknown, and it is essentially epistemically opaque (EEO) if it's impossible to know those relevant elements, those definitions come from Humphreys (2009) and are more defined in his paper where he argues that computational sciences remove humans from the center of epistemological enterprise and this removal begets novel philosophical problems. One problem is what he calls *anthropocentric predicament*, which is caused by our use of external superior, non-human, epistemic authorities in many fields, authorities that transcend our human abilities. This concept meshes with other philosophical concerns brought up in the 2nd category about how much of our humanity will have to adjust or get discarded to fit the new AI guided world.

The predicament raises 3 issues other than EO but for our scope here we limit our focus on EO and EEO, the answer by Durán and Formanek (2018) is CR which could be stated as: If a cognitive agent believing any truth-valued proposition related to the results of a computer simulation at any given time results from reliable computer simulation, then that agent's belief in that proposition at any given time is justified. In other words, if the process yields trustworthy results most of the time, then the probability that the next set of results of is trustworthy is greater than the probability that the process was unreliable and yielded trustworthy results by mere luck.

CR relies on four sources, Verification and validation methods, Robustness analysis for computer simulations, A history of (un)successful implementations and Expert knowledge. These sources vary by strength and are meant to assess the outcome to make sure that a computational process is reliable.

As we can see the answer to the risk of knowing less in this case seems to be accepting it by assessing results after the fact not understanding the process itself, this is not necessarily an issue with the concept of CR but it has implications that limit the trust in some applications, mistakes in the health sector or in military applications could be irreversible, CR is incompatible with the "Traceable" ethical principle of DoD.

At the same time if there was no more pragmatic way to peer through the Epistemic Opacity and we accept CR then sooner or later AI will fully take their roles as Guides epistemologically, no longer will we take the AI output as a 2nd opinion, but rather the AI will decide, and humans will be mere operators.

Even the sources for CR may become less reliable with time, because systems will become more complex and future experts in sensitive fields, like current chess champions for example, will start their training by listening to what the computers say and no longer through accumulated knowledge gathered when humans were the core of epistemics. At that point in the distant future the predictions of AI religions may become closer to reality.

Overlapping Risks:

Some risks result from an overlap in categories, or in sets of risks in this category, here are two examples:

Closing windows for meaningful protest or revolutions:

If we combine the risk of epistemic regression with some of the risks of AI military, we see another bleak scenario of maximizing tyranny and/or criminality.

Consider how easy it will become for governments to manipulate audiovisual media and do character assassinations of reformers or opposition party members, and how real leaks by whistleblowers could be disregarded as fake.

On the criminal side, one of the risks Morgan et al. (2020) file under the Strategy label is that of proliferation. Many AI powered devices and software is readily available for commercial use, they could easily be outfitted to become lethal, and to be used by criminals or terrorists. It could be argued that once such a thing happens both the field of AI and more civil liberties will be stifled under the pretense of counter terrorism.

Another risk comes from the Ethical and Legal label from the same report, AI may enable the establishment of surveillance states that dictators only dream of, with technologies such as facial and voice recognition, big data analysis and so on. The internet is already rife with bots manufacturing consent of the masses, while some might argue that this may also empower dissidents but the computational powers and the infrastructure that governments may use outweighs non state actors, and as history shows the state concentrates powers to stamp out dissidence much harsher than it would to fight criminals. The overlap of epistemic regression and military AI means disempowered average citizens and dissidents, coupled with AI-empowered states and cyber-criminals.

The overlap of this eventuality with the economic setbacks leads to another layer of tyranny, since tech companies are leading the charge of deploying AI there is already a sense of powerlessness on how much individuals can control their fates, a point brought by the Anti-Tech camp indirectly and by Amy Webb more directly in her book "The Big Nine", where she predicts one possible future of learned helplessness with regards to the direction tech companies take us (Berwick, 2019), and if I may add a recent example of rebranding Twitter despite the fact that none of the users asked for that. If we consider how much time users spend on social media apps we cannot understate the nuisances of such changes, but it is the least of the worries compared to more severe infringement of users' privacy and freedoms, big tech companies may collude with governments and make it impossible for dissidents to organize anywhere online, except perhaps on darker recesses of the web.

Natural Risks: damage to the environment & to human nature:

If we consider the word nature to mean both our environment and our natural selves as evolutionary products of the environment then two risks could be grouped as "Natural risks", it falls in the intersection between the 2nd category (risks to human nature as explained above) and 3rd categories (environmental damage as explained below).

AI advantages come at a cost of environmental impacts, this duality is captured by the terms "Red AI" and "Green AI" by Schwartz et al. (2020). Red AI refers to AI research that is power hungry, by sampling 60 AI papers and analyzing the computational costs, it was found that some of the large models surveyed had millions and billions of parameters, the main argument isn't that these models aren't accurate or unbeneficial, they are extremely so. But they still focus on accuracy more than efficiency, and the environmental impacts have diminishing returns in some sense. Their proposed solution is to pivot towards Green AI, which they define as AI research that doesn't sacrifice novelty

yet is mindful of the computation costs and is transparent in its reporting of power usage. Basically, when it comes to the environmental impact, there is a tradeoff between how good a model can perform and how much power it consumes, these tradeoffs according to Schwartz et al. (2020) should be studied and Red AI should be gradually turned Green through efficient use of hardware, metrics, and training methods.

The environmental risk of AI is also offset by its ability to provide unique solutions to environmental problems. However, confirming that the tradeoff is a net positive requires quantification of the power it's devouring to offer such solutions.

Summary:

This category of risks presents the most pressing issues that should be considered directly and urgently, it is the most severe set of risks, it's placed as the 3rd category because the solutions could not be adequately considered without the previous two categories.

Economic Setbacks: AI enthusiasts claim that technological advancements in the past changed the nature of work, but it did not render humans completely jobless, and so they reject that this time might be any different. I find issue in the logic that carries that statement and the contradictory statements made by the same enthusiasts about how powerful AI will be, a cursory glance at some of their arguments gives the impression that they want to have their cake and eat it too, that AI is as safe as previous technological advancements so we must not worry, at the same time we should be glad about how drastically it will improve everything, it's all pros and no cons.

The Alignment Problem: This problem could be defined as the problem of divergence between the values of the programmer and the values of the AI, it may be caused by a flaw in the reward system, or a biased data set that sets off bias in the decision-making, the definition of this problem could also be expanded to cover all risks related to AI and technology.

Military AI: There is a wide set of risks that ranges from unmanned AI weapons killing a citizen because it pegged him for a combatant, all the way up to changing the nature of warfare by nullifying nuclear weapons as deterrents. Governments are already aware of many ethical and practical issues that stem from incorporating military AI. However, as the tension rises between superpowers it would be naïve to believe that any party would give up its advantages due to these concerns, I claim that nothing short of an AI version of M.A.D would stop full blown AI capabilities from being deployed.

Epistemic Regression: As advanced technological problems may require advanced technological solutions, this makes us more and more dependent on technology, and if AI becomes powerful enough thanks to its Guide stage, then we will no longer be able to trust our senses or judgment. Even the most mundane videos and pictures we see online or the most innocuous message from a friend might no longer be considered real unless we consult another AI guide. It's only a matter of time before AI gets to its Guide stage in critical fields like the medical field and the military, furthermore, some already argue that we can rely on algorithms depending on outcomes without having to fully trace the processing.

Risk IV: Speculative Future Risks, welcoming our AGI overlords.

The risk of worrying about risks:

Gürkaynak et al. (2016) argue that fear mongering about AI risks might impede its development and humanity in the future will look back at our decisions with disdain, they start off with the first category of risks per this thesis, to give an example of how far off reality the technological predictions could be, a 1976 book called 'The World in 2010', predicted that we'd be living on three planets by now.

The main point of the article is to warn against early regulation, it proposes a threshold test to enact measures if AI becomes a threat, meanwhile it's better to allow it to develop to its full potential without any AI Winters, and since there are other existential threats that humanity faces then AI might provide the solutions.

The other main point is to consider the legal ramifications of developing AI, and how regulations that are proposed by those warning of such risks, like the open letter from Future of Life Institute and signed by Elon Musk, Stephen Hawking, and hundreds of AI researchers does. It asks why would that group of signatories or any group for that matter be given the authority to decide on regulating AI?

The final point that Gürkaynak et al. (2016) make is about the futility of regulating ASI which would be too advanced for us, we're relatively as hopeless as chimpanzees trying to regulate humans, a point that Kaplan and Haenlein (2019) take further on, stating that humans share more with chimpanzees than we will with any sentient machine, I agree and argue that therefor we can only be agnostic about ASI decisions, and this is where a lot of AI doomsayers firmly disagree.

To a certain subset of those worried about AI all the foregoing categories of risks pale in comparison to what might happen if AGI is ever developed, in fact they are quite creative when it comes to imagining scenarios of AGI brining all sorts of new risks with the ultimate one being the extinction of human beings if we don't get it right from the first time. Stephen Hawking stated that AI could stand to infinitely help us or it could be the worst event in the history of human civilization. (Kharpal, 2017)

It's no wonder that this category of risks is the one closest to fiction, after all the idea that machines will become hostile to humans is an idea that started with fictional stories not with the computer scientists. When the topic of AI or machine risks is brought up the average person thinks of movies like The Terminator where something like ASI came online and launched a nuclear war after humans tried to shut it down, or The Matrix where machines enslaved humans and used them as batteries, after humans blocked out the sky to cut off solar power that energized these sentient machines.

It is understandable that such fictional stories are no longer wild fantasies but possible scenarios, if we consider the earliest conceptualization of intelligent artificial agents in the 1st chapter such as the Phaeacians ships that can direct themselves, a mythological poem almost 3 thousand years ago became somewhat of a reality with AI in 2021 when Yara debuted ships that can load and offload its cargo, recharge itself and autonomously navigate the waters. (Klesty, 2021)

According to McCorduck (2004) the fear of intelligent machines is serious from the onset of AI, the early success in programs like Samuel's checkers program, Bernstein's chess playing program and Gelernter's geometry theory proving program made IBM's sales executives nervous that their product might scare customers, a decision was made to market computers as quick morons that are only good at following our instructions.

There are many scenarios of people warning of the risks of AI reaching or exceeding human intelligence, this category of risks refers to them but will not go into details because so far they sound more science fiction than empirical scenarios based on AI history, as we saw in chapter 2 there was not an inkling of general human level intelligence in computer games.

To give an example of other ASI scenarios there is the infamous "Roco's Basilisk", a thought experiment that was written as a blog post, picture a benevolent ASI that will punish anyone who in the present was aware of the potential of ASI but didn't aid its development. The blog was met with

ire from the website's owner who banned the topic, it caused distress for some readers as well. (Millar, 2020)

Such horrific scenario may be considered an information hazard which Bostrom (2011) defines as the risk that comes with spreading information which is in fact true but may cause or enable harm. Bostrom lays down many types but for our purposes we may say this type of information hazard is what he called Enemy Hazard under Adversarial risks section, AI to be considered a potential enemy that might become stronger and more dangerous if it obtains certain information. Bostrom did define an *Artificial Intelligence Hazard* where the "threat would derive primarily from the cognitive sophistication of the program rather than the specific properties of any actuators to which the system initially has access". (Bostrom, 2011, p. 24). This hazard is more in line with the epistemic regression problem not what is being discussed in this category.

A fictional example that comes close to the concept of information hazards comes from a novel by Chuck Palahniuk titled Lullaby, the novel has a "culling song" that kills whoever listens to it. If Roco's Basilisk is true, then we can add another plot of the Basilisk sparing those who were not aware of the risk but not those who learned about it and still chose not to aid its development, which by now includes anyone reading this thesis.

When we consider that AI learns from the data that we feed it then the Basilisk could become a self-fulfilling prophecy, in fact all the scary fictional AGI or ASI scenarios that could be thought of might be considered information hazards that bestow ideas on future AI bots.

But here we must pause and ask ourselves how deep into mere fiction are we delving? The term "Information Hazard" may sound contemporary enough to disallow us from seeing how primitive it could be, we would like to think that we have evolved and stopped believing in superstitious tales about spirits or ghouls that haunt our surroundings, yet with these notions spreading it is hard to argue that we truly progressed, especially if we revert to making certain utterings a taboo to ward off AI curses. Bostrom offers some ideas to counter these hazards, one I agree with is that we can tolerate the hazard, but the one his paper seems to be driving at is impeding certain fields.

At the same time, it would be naïve to discount all the possible risks of AGI or ASI as mere fantasies, after all what ancient civilization wrote as fiction is now a reality for us, my aim in this section is not to say that these possible scenarios are without merit. As a fiction writer myself I am fascinated by the possibilities that AI spawns. The Basilisk led me to conceive of a reverse scenario, instead of a future ASI that acts in hindsight, its ancestor may already exist or will be programmed very soon, and instead of retroactive justice it will actively develop itself by eliminating those that hinder its growth.

However, I must make a clear distinction between fiction and reality, the dystopian stories that use AGI or ASI as a trope seem to be not much more than fiction; there is no empirical data or a historical analogy to back the stories. A good example is the idea that ASI takeoff must be done right from the first time and if not, then hell breaks loose, but there is no realistic precursor to this idea, we have never dealt with an intelligence far better than ours. Nor was any AI advancement this abrupt and autonomous, usually it's steady and even tedious with a lot of human fine tuning involved.

The scenarios are helpful in investigating but they aren't an exact science, nor should they act as policy recommendations, the scenarios do not tell us something that we don't already know about how we must deal with AI if we consider all risk categories above, and even consider potential apocalyptic scenarios of ASI, it becomes obvious that there is nothing more to do other than to regulate AI, which is a conclusion one might come to without such stories. A call for an outright ban is also something that intellectuals of primitivist inclinations or extreme anti-tech views argued for regardless of ASI.

I would like to address one nihilistic argument before moving on to the final section of this category, an argument I came across while researching this topic and discussing it with computer scientists. It goes along the lines of "If AI became advanced enough and started doing evil things to humans, it is only continuing what humans are doing to each other", and indeed the problem of bias in the data

wouldn't exist if humans themselves weren't biased, however there is one element that separates all humans from machines by definition. Humans, no matter how evil, are still humanly self-preserving, even if we augment the argument by saying evil is an evolutionary necessity, and only the evil survive, it would still be a story of human survival, whereas a machine isn't hardwired to take human survival as an imperative.

The pincer of regulation and competition:

My disagreement with Andreessen (2023b) in the previous category when it comes to job loss and inequality turns into agreement here when it comes to the doom prophets, in the same article he invokes the concept of Bootleggers and Baptists, those two types are often observed when certain regulations are in place, the Baptists refer to those who genuinely believe in the importance of the regulation whereas the bootleggers benefit from the regulation in the black market. Andreessen warns of the risk of bootleggers in the form of CEOs attempting to form a cartel, and to curb competition.

But short term competition might jeopardize long term success, Kaczynski, who is the polar opposite of Andreessen, gives an example of rival kingdoms in a forest, the kingdoms that cut down trees recklessly for the war effort will have a short term advantage over the kingdoms that regulate their use of wood to keep the forest for future generations, the reckless kingdom would win the war but ultimately it would meet its demise due to ecological disaster.

If we apply this analogy to the development of AI then it's clear that it would not be sufficient for some countries to regulate this technology and subject themselves to countries that recklessly develop AI, and since the forest is the entire world then international treaties must be signed and strictly adhered to. This is most likely wishful thinking, but the time is due for humanity to start strategizing as a collective, although some would argue the weight of worldwide interconnectivity and regulation is already suffocating the freedom of everyone.

The AGI doomsayers must address the fact that harmful AI could be developed even if countries agree to be careful with its development, unlike heavy weaponry or nuclear weapons that require reactors and precious material, a few computers and access to the internet is enough. If tighter regulation and surveillance is required to make sure some fictional scenarios from happening then one can't help but see how this loops back to the risk of AI allowing governments to be more tyrannical, the bootleggers in that scenario are world governments themselves.

The point of the dialectic in this category is to present the risk that moves as a pincer with regulation and competition attack our flanks and an escape from one will force a showdown with the other. It's also worth pointing out that the tradeoff of the environmental risks falls under this pincer as well, a competition requires higher performance, but mutual longevity requires Green AI regulations.

Summary:

One can think of scenarios where the most benign machine could be misaligned enough to potentially pose an existential risk that the evilest humans can never realistically pose. This section and the chapter at large aren't meant to disregard the worries of AGI eschatology but to point out that we can't spend too much time worrying about science fiction when the scientific reality is this bleak.

I agree with Ford (2016) who argues that the focus on computers reaching AGI levels misses the point about how good computers are getting at specialized routine tasks, AI is still far from becoming AGI, yet it can already pull astonishing feats as landing aircrafts, do Wall Street trading and defeat the top champions in all the games that humans consider the most intelligent. He puts it in a more direct way in chapter 9 about Super Intelligence, where he states that computers don't need to become as smart as humans to replace them.

My main argument in this category is that the AGI alarm bell is too loud and that it drowns the more pressing and meta risks that were briefly introduced.

The ultimate challenge when it comes to a tool as powerful as AI, a tool that has the potential to become a guide and to reshape our epistemology, ethics and lives as a whole, is the challenge of balancing between unity and competition. Regulations will stifle advancements in a technology that may help us greatly, and it might put whoever self-imposes those regulations at a significant disadvantage in the face of those who are laxer with their regulations. Viewed from the other side, the competition that drives innovation and human behavior might become a race to the bottom if this technology is developed recklessly by businesses that are only looking to make profits or countries that are only trying to dominate the world.

The aim of this chapter is not to despair but to show how complex and interconnected the risks are, and to argue that any serious attempt to alleviate any danger shouldn't be considered in isolation. The risk model is meant to serve as a road map, further research should build a solution model that takes into account the pitfalls of ignoring the connections.

The following figure clarifies how the risks relate to one other. Some compromises were made for clarity's sake.

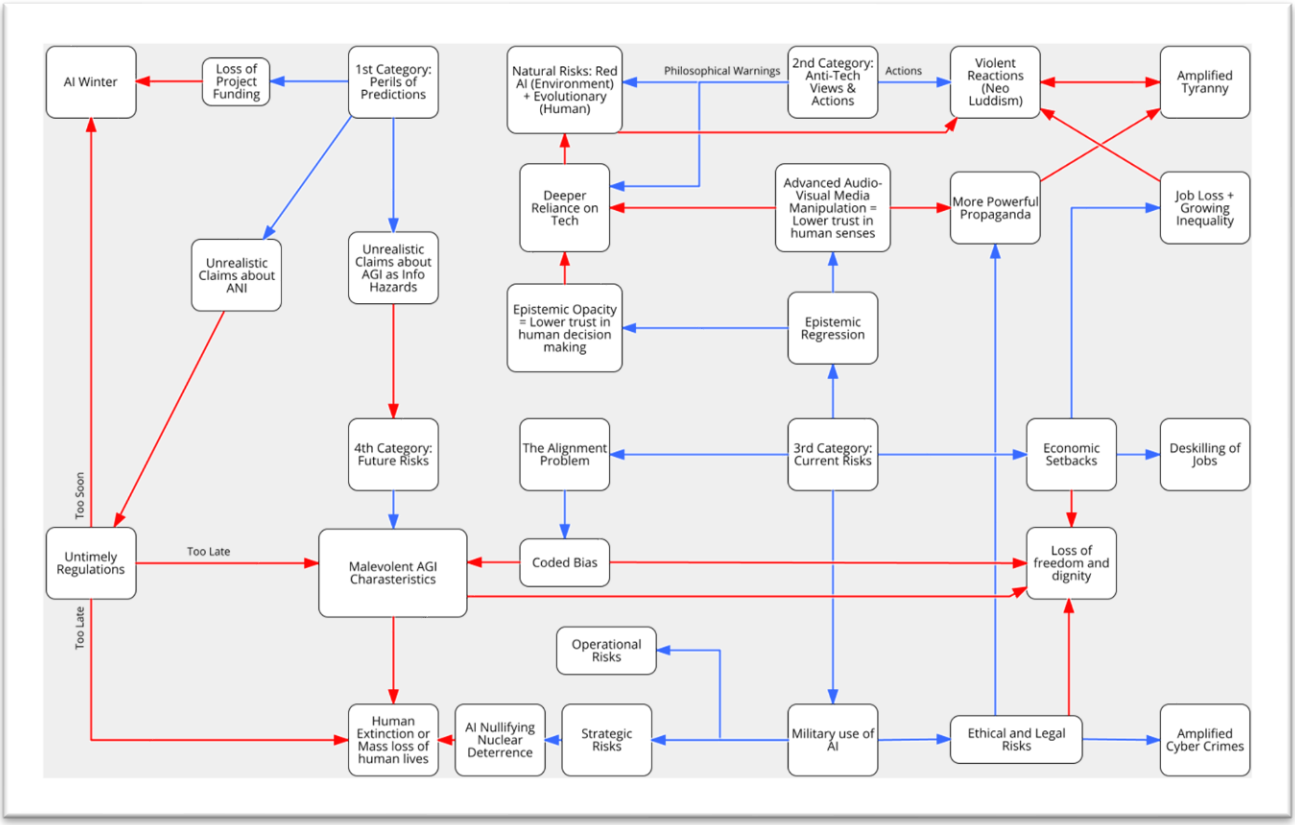

Figure 1: Risk Model. A visual representation of the relationship between AI risks as per chapter 3 categorization. The blue lines mean "includes" and the red lines mean "causes".

Conclusion:

The thesis is aimed to answer two main research questions, 1. How did AI advance in board games? and 2. What risks does AI dominance entail?

To answer the question three steps were taken in three chapters, the first chapter gives a brief overview of AI evolution through the ages to contextualize the field of AI and to prove that computers in general and AI particularly are linked with games that require strategizing, especially chess. Both Charles Babbage, who originated the idea of computers in the 19th century, and Alan Turing, who is considered the father of Artificial Intelligence, considered the application of computers in chess.

The second chapter deals with the 1st research question, four games were discussed: Backgammon, Checkers, Chess and Go. These games were selected because each represents a step in the evolution of AI. Backgammon was the first game that witnessed a win against a human champion in an exhibition match. Checkers was the first complex game that was weakly solved; the result of perfect play was concluded through computation. Chess is the most famous game that involves human champion vs machine match.

Programmers originally wanted the computer to learn by itself and not to be fed ideas by humans, in 2016 Alpha Go that learned through playing against itself won against the 9 dan Go champion Lee Sedol, the full realization of the original programmers' dream happened with Alpha Go Zero which taught itself without reference to expert matches and managed to surpass its predecessor, the remaining step is to create a more general program that can win all sorts of games, this is already happening with MuZero which starts learning without knowing the game rules.

Upon reviewing the relevant literature of these programs and others, a pattern of human-AI dynamics emerged, it could be split into 3 stages, first AI like any other technology starts as a tool, then after a certain threshold it becomes an adversary to humans unlike other tools, that is if we consider intelligence to be unique to humans, but if it's considered as any other feature such as speed, then AI is no different than modern means of transportation, the worries stem from anthropomorphizing AI.

The last stage is AI acting as a guide, even the best in their field could learn from and consult to improve. However, one of the main conclusions of this thesis after proposing this 3-stage cycle, is that the adversary stage was part of the game designs and not an inherent feature of AI. There was no autonomy in any form at any time, AI is an excellent tool and humans seem to have a hard time in comparing it to other tools because of the long held belief that intelligence is unique to humans.

The 3 stages Tool-Adversary-Guide, or T.A.G for short, could provide an analogy for the path that AI would take in any given field, if this hypothesis proves true with further research into other histories and current advancements, then computer scientists, policy makers and business owners must aim to cross the gap from Tool to Guide without any form of Adversity to humans, and to understand that adversity may manifest in different ways depending on the field.

The 3rd chapter categorizes the risks and shows how they connect to one another, to show how feedback loops in the network of risks may make the solutions harder to come by. Any attempt to solve one set of risks without understanding the interrelation may accidentally strengthen another set of risks. The thesis doesn't attempt to give solutions to those risks, as that would require more technical details and surveys that are outside the scope. But the solution to any problem starts with correctly diagnosing it, and the aim of the 3rd chapter is to do so.

After reviewing the relevant literature and the latest advancements in AI, the risks were placed into 4 categories, the term risk is not limited to the risks that AI pose but also to those that may be posed to AI. The first category deals with issues regarding predictability; AI predictions are wild sometimes and miss the mark in many ways. Failure to predict leads to overpromising and causes loss of funding in AI research.

The second category involves Anti-Tech Views, this brings attention to risks brought up by philosophers that question the trajectory of technology at large, an example is the risk of over reliance on technology and how that might significantly harm human freedom and dignity. The category also includes another risk by bringing attention to historical precedents of violent reactions to technological progress. This category seeks to fill a gap in recent literature and to question underlying assumptions about technology.

The third category includes current Risks, these risks are often associated with emerging disruptive technologies, such as Economical Setbacks due to job losses. Other unique risks to AI like The Alignment Problem which is the divergence of values between man and machine, concerns of using AI for military purposes, and finally the issue of Epistemic Regression which refers to the displacement of human senses and rationality as the main epistemic inputs.

Finally, the 4th category with regards to possible future risks, it briefly addresses the speculations about Artificial General Intelligence or Super Intelligence causing catastrophes for humans, it also addresses that such fears may be more fiction than reality and the counter view that the real risk is stopping AI research.

All in all, the review supports the conclusion that most of the negativity comes from anthropomorphizing AI and is based on scares from science fiction not the scientific reality, that is not to say AI doesn't pose serious risks, but it seems to only amplify risks that technology already poses. AI stands to reshape our lives and may be a revolution of its own, and it is crucial to steer it in the right direction, for that end we cannot ignore its history and all philosophical assumptions regarding technology, intelligence and human freedom and dignity.

References:

A. L, Samuel. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM journal of research and development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Air & Space Forces Magazine. (2023, January 20). The Near Nuclear War of 1983 | Air & Space Forces Magazine. Retrieved August 5, 2023, from https://www.airandspaceforces.com/article/the-near-nuclear-war-of-1983/

AlphaDogfight Trials Foreshadow Future of Human-Machine Symbiosis. (2020, August 26). https://www.darpa.mil. Retrieved August 6, 2023, from https://www.darpa.mil/news-events/2020-08-26

Andreessen, M. (2023b, June 6). Why AI Will Save The World. *Marc Andreessen Substack*. https://pmarca.substack.com/p/why-ai-will-save-the-world

Anslow, L. (2022). In 1903, New York Times predicted that airplanes would take 10 million years to develop. *Big Think*. Retrieved July 27, 2023, from https://bigthink.com/pessimists-archive/air-space-flight-impossible/

Armstrong, S. D., Sotala, K., & Héigeartaigh, S. Ó. (2014). The errors, insights and lessons of famous AI predictions – and what they mean for the future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 317–342. https://doi.org/10.1080/0952813x.2014.895105

Avi-Yonah, S. (2023, July 23). 'Judeo-Christian' roots will ensure U.S. military AI is used ethically, general says. Washington Post. Retrieved July 27, 2023, from https://www.washingtonpost.com/national-security/2023/07/22/air-force-general-ai-judeochristian/

Balsamo, M., Offenhartz, J., & Sisak, M. R. (2023, June 11). "Unabomber" Ted Kaczynski died by suicide in prison medical center, AP sources say | AP News. AP News. Retrieved July 31, 2023, from https://apnews.com/article/ted-kaczynski-unabomber-1197f597364b36e56bdbcaca9837bdc4

Baraniuk, C. (2022, 24 February). The cyborg chess players that can't be beaten. *BBC Future*. Retrieved March 15, 2023, from https://www.bbc.com/future/article/20151201-the-cyborg-chess-players-that-cant-be-beaten

Berliner, H. J. (1980). Backgammon Computer Program Beats World Champion. *Artificial Intelligence*. https://doi.org/10.1007/978-1-4613-8716-9_2

Berwick, I. (2019, June 21). Book review: The Big Nine by Amy Webb. Financial Times. https://www.ft.com/content/1eb9652a-22ea-11e9-b20d-5376ca5216eb

Bostrom, N. (2011). INFORMATION HAZARDS: A TYPOLOGY OF POTENTIAL HARMS FROM KNOWLEDGE. Review of Contemporary Philosophy, 10, 44–79. https://www.ceeol.com/search/article-detail?id=44170

Bridle, J. (2018, 16 July). Rise of the machines: has technology evolved beyond our control? the Guardian. Retrieved May 4, 2023, from https://www.theguardian.com/books/2018/jun/15/rise-of-the-machines-has-technology-evolved-beyond-our-control-

Chessbase (2005, 24 November). 8:4 final score for the machines – what next? *Chess News*. Retrieved March 16, 2023, from https://en.chessbase.com/post/8-4-final-score-for-the-machines-what-next-

Cheong-mo, Y. (2019, November 27). (Yonhap Interview) Go master Lee says he quits unable to win over AI Go players. https://en.yna.co.kr/index. Retrieved August 15, 2023, from https://en.yna.co.kr/view/AEN20191127004800315

Christian, B. (2021). The alignment problem: How Can Machines Learn Human Values? Atlantic Books.

Cirasella, J., & Kopec, D. (2006). The History of Computer Games. *CUNY Academic Works*.

Clarke, S., Cottier, B., Englander, A., Eth, D., Manheim, D., Martin, S. D., & Rice, I. (2022). Modeling Transformative AI Risks (MTAIR) Project -- Summary Report. *https://arxiv.org/*. https://arxiv.org/abs/2206.09360

Claude E. Shannon (1950) XXII. Programming a computer for playing chess, Philosophical Magazine Series 7, 41:314, 256-275. http://dx.doi.org/10.1080/14786445008521796

Cole, D. (2004) "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy.* Retrieved 2 April 25, 2023, from https://plato.stanford.edu/archives/win2020/entries/chinese-room/

Computer Chess Rating Lists CCRL 40/15 - Index. (z.d.). *Https://Computerchess.Org.Uk/*. Retrieved March 16, 2023, from https://computerchess.org.uk/ccrl/4040/index.html

Copeland, S. (2016, 13 January). Komodo Beats Nakamura In Final Battle. *Chess.Com*. Retrieved March 16, 2023, from https://www.chess.com/news/view/komodo-beats-nakamura-in-final-battle-1331

Coveney, P. V., & Highfield, R. (2021). When we can trust computers (and when we can't). Philosophical Transactions of the Royal Society A, 379(2197). https://doi.org/10.1098/rsta.2020.0067

Crevier, D. (1993). The Tumultuous History of the Search for Artificial Intelligence. BasicBooks.

Dewdney, A. K. (1984). Computer Recreations. *Scientific American*, *250*(5), 14–22. https://doi.org/10.1038/scientificamerican0584-14

Doggers, P. (2020, 13 April). Smerdon Beats Komodo 5-1 With Knight Odds. *Chess.Com*. Retrieved March 16, 2023, from https://www.chess.com/news/view/smerdon-beats-komodo-5-1-with-knight-odds

Durán, J. M., & Formanek, N. (2018). Grounds for trust: essential epistemic opacity and computational reliabilism. Minds and Machines, 28(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Durán, J. M., & Jongsma, K. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-106820. https://doi.org/10.1136/medethics-2020-106820

Weiss, E. F. (1992). Biographies: Eloge: Arthur Lee Samuel (1901-90). IEEE Annals of the History of Computing, 14(3), 55–69. https://doi.org/10.1109/85.150082

Ellul, J. (1964). The technological society. Vintage.

Ensmenger, N. (2012). Is chess the drosophila of artificial intelligence? A social history of an algorithm. Social Studies of Science, 42(1), 5–30. https://doi.org/10.1177/0306312711424596

European Commission - Press corner. (2023, February 21). [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_953

Ford, M. (2016). Rise of the robots: Technology and the Threat of a Jobless Future. Basic Books.

Frana, P. L., & Klein, M. J. (2021). Encyclopedia of Artificial Intelligence: The Past, Present, and Future of AI. ABC-CLIO.

Friedel, F. (2019, December 12). MuZero figures out chess, rules and all. Chess News. https://en.chessbase.com/post/muzero-figures-out-chess-rules-and-all

Funk, J., & Smith, G. (2021b, May 4). Why ambitious predictions about A.I. are always wrong. *Slate Magazine*. Retrieved July 27, 2023, from https://slate.com/technology/2021/05/artificial-intelligence-moonshots-usually-fail.html

Future of Life Institute. (2015, 28 October). Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter - Future of Life Institute. Future of Life Institute. Retrieved May 4, 2023, from https://futureoflife.org/open-letter/ai-open-letter/

Future of Life Institute. (2023, March 22). Pause Giant AI Experiments: An Open Letter - Future of Life Institute. Future of Life Institute. Retrieved May 4, 2023, from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Garvey, C. (2018). Broken Promises and Empty Threats: The Evolution of AI in the USA, 1956-1996. Technology's Stories. https://doi.org/10.15763/jou.ts.2018.03.16.02

Ghose, T. (2017, 1 February). All In: Artificial Intelligence Beats the World's Best Poker Players. *livescience.com*. Retrieved April 25, 2023, from https://www.livescience.com/57717-artificial-intelligence-wins-texas-hold-em.html

Gurkaynak, G., Yılmaz, İ., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. Computer Law & Security Review, 32(5), 749–758. https://doi.org/10.1016/j.clsr.2016.05.003

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Heidegger, M. (1977). The question concerning technology, and other essays. Facsimiles-Garl.

Hsu, F. (2002). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. Synthese 169, 615–626. https://doi.org/10.1007/s11229-008-9435-2

Hyatt, R. E., & Nelson, H. W. (1990). Chess and supercomputers: details about optimizing Cray Blitz. *Conference on High Performance Computing (Supercomputing)*, 354–363. https://doi.org/10.5555/110382.110453

IBM Corporation. (2011). IBM100 - A Computer Called Watson. *https://www.ibm.com/*. Retrieved April 25, 2023, from https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/

Kaczynski, T. J. (2016). *Anti-tech Revolution: Why and how* (1st ed.). Fitch & Madison Publishers, LLC.

Kaczynski, T. J. (2022). Technological slavery: Enhanced Edition. Fitch & Madison Publishers, LLC.

Kantayya, S. (2020). Coded Bias. Retrieved August 8, 2023, from https://www.netflix.com/title/81328723

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Karimi, F. (2023, April 29). 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. Retrieved August 8, 2023, from www.edition.cnn.com. https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html

Kharpal, A. (2017, November 6). Stephen Hawking says A.I. could be "worst event in the history of our civilization." CNBC. Retrieved August 12, 2023, from https://www.cnbc.com/2017/11/06/stephen-hawking-ai-could-be-worst-event-in-civilization.html

Kidwell, P. A. (2015). Playing Checkers with Machines—from Ajeeb to Chinook. *Information & Culture*. https://doi.org/10.7560/ic50405

Klesty, V. (2021, November 24). Yara debuts world's first autonomous electric container ship. Reuters. https://www.reuters.com/markets/europe/yara-debuts-worlds-first-autonomous-electric-container-ship-2021-11-19/

Knight, W. (2023, February 13). Eric Schmidt is building the perfect AI War-Fighting machine. WIRED. https://www.wired.com/story/eric-schmidt-is-building-the-perfect-ai-war-fighting-machine/

Kohs, *G. (2017). AlphaGo - The Movie*. Moxie Pictures, and Reel As Dirt.

Korosec, K. (2021b, February 18). TechCrunch is part of the Yahoo family of brands. https://techcrunch.com/. Retrieved August 17, 2023, from https://techcrunch.com/2021/02/18/anthony-levandowski-closes-his-church-of-ai/

Lohr, S. (2017, 3 July). Netflix Awards $1 Million Prize and Starts a New Contest. *Bits Blog*. Retrieved April 21, 2023, from https://archive.nytimes.com/bits.blogs.nytimes.com/2009/09/21/netflix-awards-1-million-prize-and-starts-a-new-contest/

Madrigal, A. C. (2017, 19 July). How Checkers Was Solved. The Atlantic. Retrieved May 5, 2023, from https://www.theatlantic.com/technology/archive/2017/07/marion-tinsley-checkers/534111/

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

McArthur, N. (pre-print). AI Worship as a New Form of Religion. https://philpapers.org/. Retrieved August 8, 2023, from https://philpapers.org/rec/MCAAWA

McCarthy, J. (2007). What is Artificial Intelligence? http://www-formal.stanford.edu/jmc/. https://www-formal.stanford.edu/jmc/whatisai.pdf

McClain, D. L. (2006, 5 December). Once Again, Machine Beats Human Champion at Chess. *The New York Times*. Retrieved March 18, 2023, from https://www.nytimes.com/2006/12/05/crosswords/chess/05cnd-chess.html

McCorduck, P. (2004). *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*. CRC Press.

Mendez, M., II. (2023, 20 April). The Drake AI Song Is Just the Tip of the Iceberg. *Time*. Retrieved April 21, 2023, from https://time.com/6273529/drake-the-weeknd-ai-song/

Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. Nature Machine Intelligence, 5(4), 333–334. https://doi.org/10.1038/s42256-023-00644-2

Millar, I. (2020). The Psychoanalysis of Artificial Intelligence [PhD dissertation]. Kingston School of Art. Retrieved August 7, 2023, from https://eprints.kingston.ac.uk/id/eprint/49043/1/Millar-I-49043.pdf

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). Military applications of artificial intelligence: ethical concerns in an uncertain world.

Muggleton, S. (2014). Alan Turing and the development of Artificial Intelligence. *Ai Communications*, *27*(1), 3–10. https://doi.org/10.3233/aic-130579

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M., & Floridi, L. (2023). Taking AI risks seriously: a new assessment model for the AI Act. AI & Society. https://doi.org/10.1007/s00146-023-01723-z

Piccinini, G. Turing's Rules for the Imitation Game. Minds and Machines 10, 573–582 (2000). https://doi.org/10.1023/A:1011246220923

Piesing, M. (2013, March 30). Predicting the future of artificial intelligence has always been a fool's game. *Wired.com*. Retrieved July 28, 2023, from https://www.wired.co.uk/article/predicting-artificial-intelligence

REAIM Call to Action. (2023, February 16). [Press release]. https://www.government.nl/ministries/ministry-of-foreign-affairs/documents/publications/2023/02/16/reaim-2023-call-to-action

Religious leaders call for a ban on killer robots – PAX. (2014, November 12). https://paxforpeace.nl/. Retrieved August 8, 2023, from https://paxforpeace.nl/news/religious-leaders-call-for-a-ban-on-killer-robots/

Rerun Productions. (1996). The Treachery Of Technology. YouTube. https://youtu.be/BOCtu-rXfPk?t=409

Road accidents | Statbel. (2023, June 15). https://statbel.fgov.be/en/themes/mobility/traffic/road-accidents

Rodgers, J. (2023, 1 May). Ding Liren Wins 2023 FIDE World Championship In Rapid Tiebreaks. Chess.com. Retrieved May 4, 2023, from https://www.chess.com/news/view/fide-world-chess-championship-2023-tiebreak-ding-liren

Rosling, H., Rosling, O., & Rönnlund, A. R. (2018). Factfulness: The Ten Reason We're Wrong About the World--And Why Things Are Better Than You Think.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sale, K. (1996). *Rebels Against The Future: The Luddites And Their War On The Industrial Revolution: Lessons For The Computer Age*. Da Capo Press, Incorporated.

Schaeffer, J., Culberson, J., Treloar, N., Knight, B., Lu, P., & Szafron, D. (1991). Reviving the game of checkers. Department of Computing Science, University of Alberta

Schaeffer, J. (1997). *One Jump Ahead: Challenging Human Supremacy in Checkers*. Springer Science & Business Media.

Schaeffer, J., Burch, N., Björnsson, Y., Kishimoto, A., Müller, M., Lake, R. W., Lu, P., & Sutphen, S. (2007). Checkers Is Solved. *Science*, *317*(5844), 1518–1522. https://doi.org/10.1126/science.1144079

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Searle, J. R. (1984). *Minds, Brains and Science*. Harvard University Press.

Sentient Digital, Inc. (2023, January 31). Military Applications of AI in 2023- Sentient Digital, Inc. https://sdi.ai/blog/the-most-useful-military-applications-of-ai/

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L. R., Lai, M., Bolton, A., Chen, Y., Lillicrap, T. P., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T. P., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Simos, M., Konstantis, K., Sakalis, K., & Tympas, A. (2022). "AI CAN BE ANALOGOUS TO STEAM POWER" or From the "Post-Industrial Society" To the "Fourth Industrial Revolution": An Intellectual History of Artificial Intelligence. ICON: Journal of the International Committee for the History of Technology, 27(1), 97–116. https://www.researchgate.net/publication/362231183_AI_Can_Be_Analogous_to_Steam_Power_or_From_the_Post
Industrial_Society_to_the_Fourth_Industrial_Revolution_An_Intellectual_History_of_Artificial_Intelligence

Skrbina, D. (2014). The metaphysics of technology. In Routledge eBooks. https://doi.org/10.4324/9781315879581

Somerville, P. D. H. (2018, June 2). Google to scrub U.S. military deal protested by employees - source. U.S. Retrieved August 7, 2023, from https://www.reuters.com/article/uk-alphabet-defense-idUKKCN1IX5YC

Strickland, E. (2021, September 9). How IBM Watson overpromised and underdelivered on AI health care. *IEEE Spectrum*. Retrieved July 27, 2023, from https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Taylor, J., & Hern, A. (2023, 2 May). 'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation. the Guardian. Retrieved May 4, 2023, from https://www.theguardian.com/technology/2023/may/02/geoffrey-hinton-godfather-of-ai-quits-google-warns-dangers-of-machine-learning

Tesauro, G. (1989). Neurogammon Wins Computer Olympiad. *Neural Computation*, *1*(3), 321–323. https://doi.org/10.1162/neco.1989.1.3.321

Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of The ACM*, *38*(3), 58–68. https://doi.org/10.1145/203330.203343

The First "Advanced" or "Freestyle" or "Centaur" Human & Computer Chess Event : History of Information. (z.d.). *Www.HistoryofInformation.Com*. Retrieved March 15, 2023, from https://www.historyofinformation.com/detail.php?id=4259

Thompson, K. (1982). Computer Chess Strength. *Elsevier eBooks*, 55–56. https://doi.org/10.1016/b978-0-08-026898-9.50008-5

Toews, R. (2022, December 15). What We Got Right And Wrong In Our 2022 AI Predictions. *Forbes*. https://www.forbes.com/sites/robtoews/2022/12/15/what-we-got-right-and-wrong-in-our-2022-ai-predictions/?sh=7b8497ed1654

Traiger, S. (2003). Making the Right Identification in the Turing Test. Studies in cognitive systems, 99–110. https://doi.org/10.1007/978-94-010-0105-2_4

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/lix.236.433

U.S. Department of Defense. (2020, February 25). DOD adopts 5 Principles of Artificial Intelligence Ethics. https://www.defense.gov/News/News-Stories/article/article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/

Verma, P. (2023, March 27). AI can draw hands now. That's bad news for deep-fakes. *Washington Post*. Retrieved July 21, 2023, from https://www.washingtonpost.com/technology/2023/03/26/ai-generated-hands-midjourney/

Watercutter, A. (2020, 28 October). Why "The Queen's Gambit" Is the No. 1 Netflix Show Right Now. *WIRED*. Retrieved March 16, 2023, from https://www.wired.com/story/the-queens-gambit-netflix-chess-addiction/

Wessling, B. (2022, June 24). A decade after acquiring Kiva, Amazon unveils its first AMR. The Robot Report. Retrieved August 5, 2023, from https://www.therobotreport.com/a-decade-after-acquiring-kiva-amazon-unveils-its-first-amr/

Williams, A. (2017). History of Digital Games. *Routledge eBooks*. https://doi.org/10.1201/9781315715377

Yong, E. (2019, October 2). Brain Simulation Promised a Decade Ago Hasn't Succeeded. *The Atlantic*. Retrieved July 27, 2023, from https://www.theatlantic.com/science/archive/2019/07/ten-years-human-brain-project-simulation-markram-ted-talk/594493/

Zheng, F., Liu, Z., Liu, T., Hung, C., Xiao, J., & Feng, G. (2021). Facial expression GAN for voice-driven face generation. The Visual Computer, 38(3), 1151–1164. https://doi.org/10.1007/s00371-021-02074-w

Zobrist, A. L. (1969). A model of visual organization for the game of GO. *National Computer Conference*. https://doi.org/10.1145/1476793.1476819

Appendices:

Appendix 1: T.A.G Stages Summary.

Stage\Game	Backgammon	Checkers	Chess	Go
Tool	Hans Berliner started programming *BKG.* (1974) Gerald Tesauro started working on *Neurogammon.* (1989)	Christopher Strachey and Arthur Lee Samuel started working on Checkers programs. (1952, 1954) IBM hosted the world championships on the condition that Samuel's program would play against Walter Hellman and Derek Oldbury, it lost all matches but won IBM 15 stock points. (1966) Jonathan Schaeffer started working on *Chinook.* (1989)	Leonardo Torres y Quevedo designed "*el Ajedrecista*", a machine that played one Chess ending. (1912) Alan Turing wrote *Turochamp*, a chess program code, before computers were able to execute it. (1948) Claude Shannon wrote the first paper on chess programming foundations. (1950) Feng Hsiung Hsu started working on *Chiptest.* (1985)	Albert L. Zobrist worked on the earliest Go program. It was only able to beat players with less than 20 games experience. (1969) DeepMind started AlphaGo research project. (2014)

Adversary	*BKG 9.8* won against Luigi Villa in an exhibition match. (1975) *Neurogammon* lost to Ossi Weiner but won against five Backgammon programs in the Computer Olympiad. (1989) The 2nd version of *TD-Gammon* played 38 exhibition games against top human players and had a net loss of 7 points. (1992)	Program *Colossus Draughts* was the first checkers engine to win a human tournament. (1990) *Chinook* lost against Marion Tinsley in their first and second encounters. (1990, 1992) Chinook Draw with Tinsley. (1994) *Chinook* won against Lafferty in World Man-Machine Championship and defended its title as a champion after Tinsley's passing. (1995)	Ken Thompson's *Belle* reached the chess master rating of 2000. (1983) Hans Berliner's *HiTech* reached the 2400 rating. (1988) Deep Blue Jr. lost against Garry Kasparov. (1996) Deep Blue won against Garry Kasparov. (1997) Vladimir Kramnik lost to Deep Fritz. (2006)	AlphaGo won against Fan Hui. (2015) AlphaGo won against Lee Sedol (2016) HanDol, a Korean Go AI program, won with two pieces handicap against Sedol. (2019)
Guide	TD-Gammon used an opening "splitting" which was deemed inferior by top players compared to "slotting", Bill Robertie wrote an article that shows the program analysis proves that it's the other way around, later "splitting" took over as a much more common opening. (1992)	Tinsley and Elbert Lowder were playing a match that was taking too much time, they decided to have *Chinook* adjudicate. (1994) Schaeffer and his team weakly solved Checkers. (2007)	Centaur chess match between Kasparov with the aid of Fritz 5 and Topalov with the aid of ChessBase 7.0 was held, the result was a draw. (1998) Human players currently turn to chess programs for practice and analysis. Champions already started to learn by playing against computers but the 1997 match cemented chess engines' superiority.	AlphaGo Zero managed to beat other AlphaGo Variations. (2017) AlphaZero proved superior to other game engines in chess (Stockfish), Shogi (Elmo), and Go (AlphaGo Zero). (2017) MuZero, which starts learning from scratch, matched AlphaZero's performance in all 3 games and can play more than 50 Atari games. (2019)